Bioinformatics approaches for functional predictions in diverse informatics environments.

Paula Maria Moolhuijzen, BSc

This thesis is presented for the degree of Doctor of Philosophy of Murdoch University

2011

Declaration

I declare that this thesis is my own account of my research and contains as its main content work, which has not previously been submitted for a degree at any tertiary education institution.

Signature:

Name: Paula Moolhuijzen

Date: 4th October 2011

Abstract

Bioinformatics is the scientific discipline that collates, integrates and analyses data and information sets for the life sciences. Critically important in agricultural and biomedical fields, there is a pressing need to integrate large and diverse data sets into biologically significant information. This places major challenges on research strategies and resources (data repositories, computer infrastructure and software) required to integrate relevant data and analysis workflows. These challenges include:

- The construction of processes to integrate data from disparate and diverse resources and legacy systems that have variable data formats, qualities, availability and accessibility constraints.
- Substantially contributing to hypothesis driven research for biologically significant information.

The hypothesis proposed in this thesis is that in organisms from divergent origins, with differing data availability and analysis resources, *in silico* approaches can identify genomic targets in a range of disease systems. The particular aims were to:

1. Overcome data constraints that impact analysis of different organisms.
2. Make functional genomic predictions in diverse biological systems.
3. Identify specific genomic targets for diagnostics and therapeutics in diverse disease mechanisms.

In order to test the hypothesis three case studies in human cancer, pathogenic bacteria, and parasitic arthropod were selected, the results are as follows.

In case study 1 sequence information was integrated to make novel predictions, and generate novel findings for the role of the Alu repeat element in cancer. An under representation of Alu was found in cancerous transcript and most noncancerous Alu transcript found were of an unknown function. These findings led to an Alu-mediated siRNA model for the down regulation of Alu containing mRNA in cancer.

Case study 2, comparative genomic analyses identified venereal diagnostic targets that discriminated *Campylobacter fetus* subspecies *venerealis* from other *Campylobacter* species and subspecies. Plasmid borne virulence Type IV secretory pathway genes specificity however varied for biovars, compromising their use for diagnostics. These findings resulted in the targeted sequencing of *Campylobacter fetus* subspecies *venerealis* biovar genomes.

Case study 3, in cattle tick ectoparasite (*Rhipicephalus microplus*), a large highly complex and under researched genome, transcript sequence was analysed and tick vaccination targets identified. These vaccine candidates successfully imparted immunity in the bovine host. The developed high throughput vaccine target identification system is now being applied to other disease systems.

Through the shared bioinformatics approaches, novel functional targets and models in disease were determined. This thesis has developed and demonstrated *in silico* approaches for:

1. The collation, annotation and integration of data from divergent organisms with variable data constraints.
2. Novel functional predictions in diverse biological systems.
3. Novel vaccine and diagnostic candidate identification, in diverse disease mechanisms, substantially contributing to hypothesis driven research.

Table of Contents

Acknowledgements

This thesis would not have been possible without the support of many good people. Thank you to my supervisors Professor Rudi Appels and Professor Matthew Bellgard for the opportunity to under take this thesis, and for their expert guidance and support. To the team at the Centre for Comparative Genomics, especially Mr David Schibeci, Mr Mark O'Shea, Dr Roberto Barrero and Mr Adam Hunter, thank you for your expert help and support. I would also like to acknowledge key collaborators at the DEEPI in Queensland and the BeefCRC, especially Dr Ala Lew-Tabor, Dr Manuel Rodreguez-Valle, Dr Felix Guerrero (ARS-USDA) and Dr Jessica Morgan. Last but not least, my deepest gratitude to my family and partner for their support and encouragement.

Abbreviations

Ab - Antibodies

Ag - Antigens

API - Application programming interface

BES – BAC End Sequence

BLAST - Basic Local Alignment Search Tool

BmiGI – *Boophilus microplus* Gene Index

CCG - Center for Comparative Genomics

COG - Clusters of Orthologous Groups

cDNA - cloned DNA

DDBJ - DNA Data Bank of Japan

DE - Differentially expressed

DFCI - Dana Faber Cancer Institute

DNA - Deoxyribonucleic acid

DPI - Department of Primary Industries, QLD

dsRNA - double-stranded RNA

EMBL - European Molecular Biology Laboratory

EBI - European Bioinformatics Institute

EST - Expressed sequence tag

FB - Fold Back

GI - Gene Index / Genomic Islands

GIRI – Genetic Information Research Institute

GO - Gene Ontology

GWAS - Genome-Wide Association Studies

HDI - Histone Deacetylase Inhibition

HMM - Hidden Markov Model

HPLC - High-Performance Liquid Chromatography

HR – Highly Repetitive

HTP – High ThroughPut

IGP – *Ixodes scapularis* Genome Project

IS - Insertion Sequence

JBIRC - Japan biological Information Research Center

KOG - Eukaryote Clusters of Orthologous Groups

LSU – Long SubUnit

LTR - Long Terminal Repeat

MGE - Mobile Genetic Elements

MR – Moderately Repetitive

miRNA - micro RNA

mRNA - messenger RNA

MSA - Multiple sequence alignment

NCBI - National Center for Biotechnology Information

ncRNA – non-protein coding RNA

NIH - National Institute of Health

NLM - National Library of Medicine

NMD - Nonsense Mediated Decay

OIE – Office International des Epizooties

ORF - Open Reading Frame

PAI – Pathogenic Islands

PANTHER - Protein ANalsis THrough Evolutionary Relationships

PCR - Polymerase Chain Reaction

pFAM - protein FAMily

PFGE – Pulse Field Gel Electrophoresis

PID - Percent IDentity

qRT-PCR - quantitative Real-Time PCR

QTL - Quantitative Trait Loci

RNA - RiboNucleic Acid

RNAi - RNA interference

SINE – Short Interspersed repetitive Element

siRNA - small interfering RNA

SNP - Single Nucleotide Polymorphism

SSH – Suppressive Subtractive Hybridization

SSU – Small SubUnit

SQL - Simple Query Language

TE - Transposable Element

TC - Tentative Consensus sequences

TGI -TIGR Gene Index

UNSAM - La Universidad Nacional de San Martín

UPM - Universal Primer Mix

UTR - Un-Translated Region

VF – Virulence Factors

VI – Vaccine Identification

Publications associated with this PhD thesis

1. Moolhuijzen P, Kulski JK, Dunn DS, Schibeci D, Barrero R, Gojobori T, Bellgard M: The transcript repeat element: the human Alu sequence as a component of gene networks influencing cancer. Funct Integr Genomics 2010, 10(3):307-319.
2. Moolhuijzen PM, Lew-Tabor AE, Wlodek BM, Aguero FG, Comerci DJ, Ugalde RA, Sanchez DO, Appels R, Bellgard M: Genomic analysis of *Campylobacter fetus* subspecies: identification of candidate virulence determinants and diagnostic assay targets. BMC Microbiol 2009, 9:86.
3. Lew AE, Guo S-Y, Venus B, Moolhuijzen P, Sanchez D, Trott D, Burrell P, Wlodek B, Bellgard M: Comparative genome analysis applied to develop novel PCR assays to characterise and identify *Campylobacter fetus* subsp. *venerealis* isolates. Zoonoses and Public Health 2007 54(Supplement 1):154.
4. Moolhuijzen P, Lew-Tabor A, Morgan ATJ, Rodriguez Valle M, Peterson GD, Dowd S. E, Guerrero F, Bellgard M, Appels R: The complexity of *Rhipicephalus* (*Boophilus*) *microplus* genome characterised through detailed analysis of two BAC clones. BMC Research Notes 2011, 22;4:254.
5. Bellgard MI, Moolhuijzen PM, Guerrero F.D., Appels R, Schibeci D, Rodriguez-Valle M, Barrero R, Hunter A, Lew-Tabor AE: CattleTickBase: Internet-based analysis tools and bioinformatics repository of available

genomics resources for *Rhipicephalus* (*Boophilus*) *microplus*. International Journal for Parasitology 2011, In review.

6. Guerrero FD, Moolhuijzen PM, Peterson DG, Bidwell S, Caler E, Appels R, Bellgard M, Nene VM, Djikeng A: Reassociation kinetics-based approach for partial genome sequencing of the cattle tick, *Rhipicephalus* (*Boophilus*) *microplus*. BMC Genomics 2010, 11:374.
7. Lew-Tabor AE, Moolhuijzen PM, Vance ME, Kurscheid S, Valle MR, Jarrett S, Minchin CM, Jackson LA, Jonsson NN, Bellgard MI et al: Suppressive subtractive hybridization analysis of *Rhipicephalus* (*Boophilus*) *microplus* larval and adult transcript expression during attachment and feeding. Vet Parasitol 2009, 167(2-4):304-320.

1 Chapter One – Introduction

1.1 Thesis structure

This thesis is set out in six chapters, with this chapter (Chapter 1) as an overview of the thesis structure, background, significance of study and aims. Chapter 2 is the literature review of bioinformatics analysis 'state of play' for resources and an introduction to three case studies for hypothesis driven research. Chapters 3-5 are the three bioinformatics case studies, human Alu repeat element in cancer, diagnostic target identification in *Campylobacter* and tick vaccine target identification for bovine. Chapter 6 is the concluding discussion on scientific contributions and future work.

1.2 Background and significance of study

The abundance of data associated with bioinformatics is increasing at an exponential rate, this is due to the increased number of sequencing projects and the technological and economical break throughs of next generation high through put sequencing. The advent of high performance computing capability has been essential for these advances. This data is critically important in biological research (agricultural and biomedical) fields, with a pressing need to integrate large and diverse data sets into biologically significant information.

Hypothesis driven analysis is an objective approach where a hypothesis must be clearly stated, analyses are planned prior, and the data must be collected in a repeatable manner. This is in contrast to the more subjective data driven analysis where analyses are *post-hoc* from data mining. Hypothesis driven approaches required for data and analysis workflow integration are a major bottleneck, especially when data and information repositories and resources are disparate and have variable data formats, qualities, availability and accessibility. Bioinformatics has many data and analysis constraints, as well as integration challenges that impact on the predictive power of *in silico* analyses, these include:

1. The lack of availability of an organism sequence data and the associated sequence information (metadata). Different organisms can be represented by only a few sequences through to whole genome projects, such as is found in the well-researched human and model species. Further scenarios are that a target species may have only a distant reference sequenced genome or none to make comparative analyses and prediction models.
2. The quality of the sequence data related to the stage of sequencing projects.
    - In different stages of an assembly project, sequence may be available as unassembled reads, assembled contiguous sequences, scaffolds with sequence gaps or completed chromosome sequences. The incomplete assembly of sequence impacts predictive powers for identifying genomic features.
    - The quality of the sequence information depends on the amount and

type of annotation and curation effort. While some sequences may be represented by variable information quality, a reference data set is expected to have high quality annotation.

3. The amount of known biological information associated with an organism. Certain organisms may be limited by the lack of biological research that has gone before.
4. Availability and access to analysis tools to process and analyse large sets of sequence data and information.
   - The ability to access and share data can be resource dependent. Resources for special research interests can develop in an uncoordinated manner, using heterogeneous database management systems and data structures.
   - Available tools and resources are too specialised for an organism, lacking robustness and flexibility.
5. *In silico* methods and approaches are poorly defined, as can be found in new emerging platform technologies (Proteomics, Metabolomics and Fluxomics).

All these constraints and challenges can play a role and limit the effective identification of functional targets in disease, especially in non-model organisms. The development of systems centred on disease have to date focused on human and model organisms (rat and mouse), and as a result targeting candidate disease features *in silico* in non-model organisms is difficult.

It is the study of these data and resource integration challenges that forms the basis of this PhD thesis. Through developed bioinformatics approaches, functional predictions based on comparative analyses and predictive modelling tools are made for the genome sequence of simple pathogens (bacteria) through to complex eukaryotes (arthropod parasite and human), to identify unknown functional and structural genomic elements that play a role in disease. In developing *in silico* approaches to identify candidate genomic disease targets for diagnostics and therapeutics, this thesis examined three diverse disease mechanisms for organisms of diverse origin and different data availability (Table 1.1). In case study 1, the well studied human transcriptome is analysed to determine from the distribution of Alu element the possible role of Alu in cancerous organs; Case study 2 investigated venereal diagnostic targets for pathogenic *Campylobacter fetus* sub species based on a close reference genome; Case study 3, in a large highly complex, under researched and unsequenced cattle tick (ectoparasite) genome (*Rhipicephalus microplus*), tick candidate vaccination targets are identified.

Case study organism	Sequence reference	Disease mechanism	Research level (sequence and biological)
1. Human	Human transcriptome and genome	Alu repeat in Cancer	High
2. Bacteria	Close subspecies genome	Pathogen diagnostics	Low-medium on comparative bacteria
3. Tick	No genome available	Ectoparasite vaccine	Low on comparative ticks

Table 1.1 Case study organisms, disease mechanisms and reference data availability

The thesis aims and case studies are further described in the next section.

1.3 The aims of this thesis

In summary the key major outstanding bioinformatics challenges from this chapter are:

- The refining of processes for data integration from disparate resources.
- Integrating data from diverse resources that have variable data formats, qualities, availability and accessibility.
- Substantially contributing to hypothesis driven research.

The particular aims of this thesis are then to develop *in silico* strategies for:

1. The collation, annotation and integration of diverse data types. To overcome data accessibility, availability, format and quality variability that impact data analyses (integration and mining) in diverse organisms.
2. The identification of functional genomic targets in diverse biological systems with diverse disease mechanisms. To make functional predictions based on comparative analyses and prediction tools.
3. High through put vaccine and diagnostic candidate identification. To identify specific genomic pathogenic and disease targets in diverse disease mechanisms.

The three overall thesis aims (above) and the corresponding chapter topics (below) are shown in Figure 1.1.

In case study 1 (Chapter 3), the human transcriptome sequence, associated sequence information and repetitive element data are integrated, for the statistical analysis of the human transposable element Alu in organ cancers, to determine the role of human Alu element in cancer.

In case study 2 (Chapter 4), a partially sequenced pathogen is assembled and annotated for comparative genomic analysis to predict pathogen subspecies diagnostic targets.

In case study 3 (Chapter 5), an under studied partial parasite transcriptome is annotated and *de novo* genomic features predicted to identify tick candidate targets, for a bovine vaccine that is protective against tick parasites.

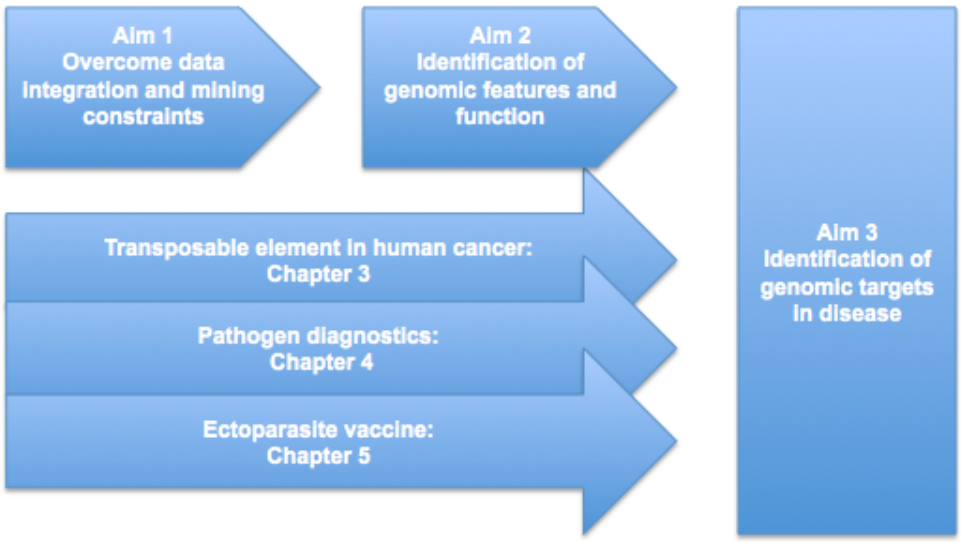


Figure 1.1 Thesis aims and case study chapters

Detail of bioinformatics approaches, thesis aims and the three case studies can be found in Table 1.2, which maps thesis aims (column 1) to the thesis approaches (column 2) and the three diverse case studies (column 3-5). To develop strategies to access data for analysis, diverse genomes selected had variable genomic sequence and reference availability and complexity (point 1 in Table 1.2). The identification of structure and functional features are explored by comparative analyses and *de novo* prediction methods. In all three cases comparative analyses were developed for making functional predictions (point 2 in Table 1.2). To develop bioinformatics approaches for identifying specific disease related genomic targets (point 3 in Table 1.2) each case study has different disease mechanisms.

Table 1.2 Table of Bioinformatics approaches, thesis aims and the three case studies.

Thesis aims	Bioinformatics approach	Case study 1 Alu element in Cancer	Case study 2 Pathogen diagnostics	Case study 3 Parasite vaccine
1. Collate, annotate and integrate diverse data types	Develop strategies to overcome data availability, format, quality and accessibility constraints	Filter heterogeneous sequence source and chart metadata to produce a data set for statistical analysis	Comparative assembly of a partial genome to a close reference genome to produce a pseudomolecule for comparative analysis	*De novo* genome assembly in the absence of a close reference genome.
2. Predict gene functions	Develop strategies for functional predictions based on comparative analyses and data integration	Chart repeat element to sequence, organ, cancer and functional annotation	Gene prediction and functional annotation. Comparative species analysis to identify sequence specificity.	*De novo* functional and structural gene predictions for comparative analyses. Cluster analysis of available transcript for comparative analysis.
3. Identify genomic targets in disease	Develop strategies to identify specific disease targets	Statistical analysis of Alu content in cancer	Subspecies specific virulence genes for diagnostics	Selection of arthropod and mammalian sequence variance and vaccine candidates

2 Chapter Two - Literature Review

2.1 General introduction

Research and development in the systems biology of plants, animals and infectious disease (host-pathogen/parasite-environment), is increasingly dependent on integrating data from diverse and dynamic sources [1]. In current bioinformatics, data integration is significantly more challenging as the volume, rate of production, and complexity of data types in bioinformatics increases [1].

Bioinformatics in the context of this thesis encompasses the tools (software and algorithms) and techniques used for the analysis of molecular biological data. Relying on computer science hardware to store data, and networks to communicate the results [2]. Often in bioinformatics a 'one-size-fits-all' program does not exist [3]. Knowledge is required about the different analysis steps in a given application and how software applications operate at each step [3]. In the construction of bioinformatics workflows the choice of software selected can influence the amount of biologically useful information that can be extracted from the data (data mining) [4].

The challenge for data integration is how to work across multiple database resources simultaneously; to integrate structured data, such as that found in relational databases, with unstructured data, such as literature; and to finally make data and information accessible in a usable format to life sciences researchers [1]. As these themes occur throughout bioinformatics, the focus of this thesis is on the particular data integration and mining challenges in three case studies. These case

studies investigate the identification of functional genomics target in disease to discover and develop new vaccines/therapeutics and diagnostics. Figure 2.1 gives an overview of thesis review topics covered in this chapter (Chapter 2).

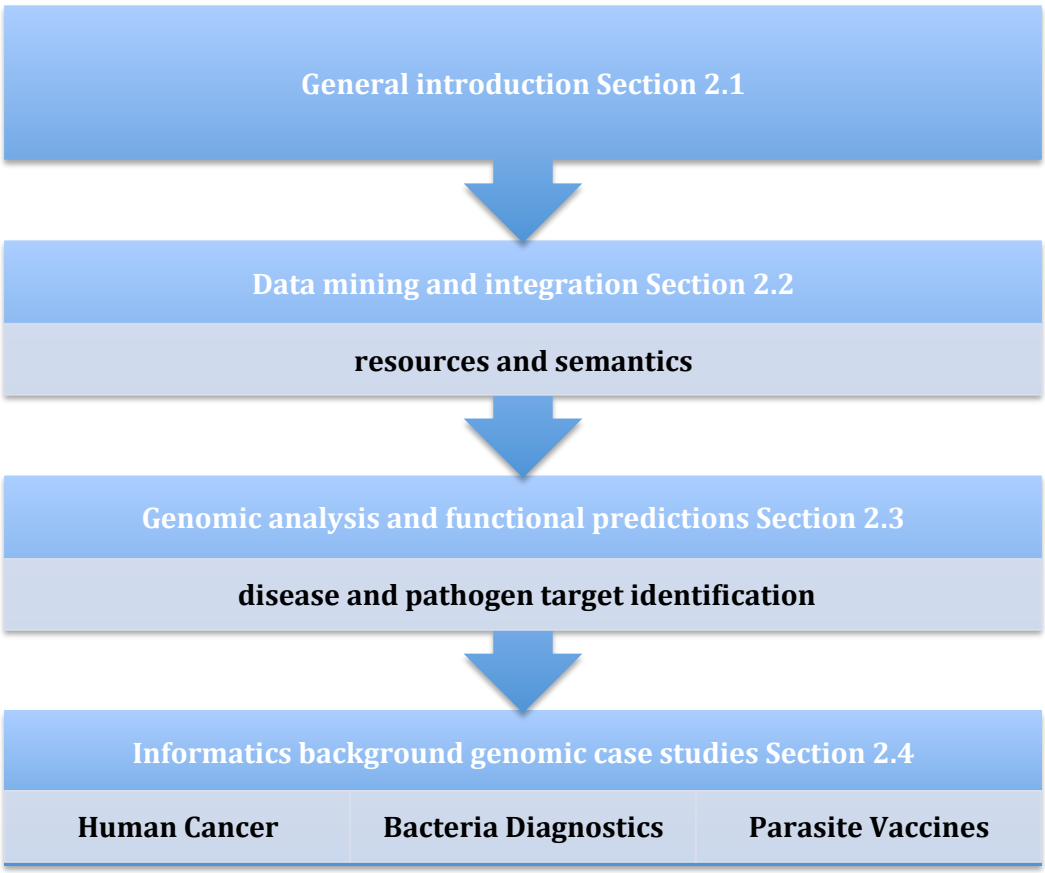


Figure 2.1 Chapter 2 topics overview

2.2 Data mining and integration

It is recognised that the ability to undertake data mining and computational analysis on a wide variety of data sources and data types has the potential to add enormous power and value to experimental analysis [5].

Some limitations however exist to access and share data. For instance, many resources (created for special research interests) develop in an uncoordinated manner, using heterogeneous database management systems and data structures. The information sets and the semantics of data can differ and integrating information from these disparate sites is difficult [6].

The ever increasing gap between next generation sequencing technology output and the ability to process and analyze the resulting sequence data [4], can delay the extraction of, and access to, important biological information. The processing and analysis of this data can be computationally intense and requires information flows to be defined in order to select the appropriate bioinformatics tools for effective analysis and data mining [3].

Suitable access to the whole human genome sequence and it's associated annotation for data mining was recognized as a key issue by the Human Genome Consortium, and a user's guide to three major human genome browsers was published [7]. Internet-based genome browser data and associated analyses were expected to expedite biological knowledge discovery. The three major genome browsers developed were the University of California at Santa Cruz (UCSC) Genome Browser [8], the National Center for Biotechnology Information (NCBI) MapViewer [9], and the Ensembl [10] Genome Browser which was developed in a joint project between the European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI) and the Sanger Institute [11, 12]. This

visualization software, designed for browsing was limited for downstream analysis and data mining, where queries can become complex and require additional information sourced elsewhere “combinational in nature” [13].

As data and resources are often decentralised, the need to standardize methodologies to facilitate information exchange and access analytical resources has been widely recognised [14]. Data integration requires standardized data models and formats, that come from a sound and thorough knowledge of the domain, with well-defined and properly managed information and data (semantics) systems, procedures and processes [6].

Romano in 2008 noted that systems are required to be flexible, expandable and adaptable to the evolving nature of biological theories and discoveries. Automated data access and analysis can only manage and control the disparate data sources as long as the correct information semantics and ontology is used [6].

2.2.1 Semantics and ontology

Integration of different platform data (Genomics, Proteomics, Transcriptomics and Metabolomics) can be limited due to lack of interoperability among life sciences resources, a perennial issue in bioinformatics [15]. Di Bernardo in 2008 noted that the global acceptance and application of standards for naming, representing, describing and accessing biological information is currently not fully adopted [16]. Important initiatives provide platforms in this area of data integration.

(1) The W3C-led Semantic Web initiative has standards and technologies to create a global Semantic Web for the Life Sciences (SWLS) [15]. Interoperability problems can be solved programmatically by annotating services and their interfaces with semantic information [16-19].

(2) Open Source research projects such as BioMOBY [20], aim to generate the architecture for the distribution of biological data through web services. Here data and resources are disparate and only the interaction instructions (object-driven registry query system with object and service ontology's) are registered in a central location. Knowledge is central and data is maintained at its primary source without a need for warehousing [21]. Objects are lightweight XML, and the query and the response transaction, simple object access protocol (SOAP) [14].

(3) Current biological ontology, structured and controlled vocabularies, are being developed and curated by a consortium [22], where Gene Ontology's (GO) terms are develop and maintained for gene products and sequences [23].

SWLS platforms provide interoperable access to thousands of bioinformatics resources, making the selection of services to build workflows at times unmanageable for users [14, 16]. A workflow assembly client can reduce the number of choices that users have to make by restricting the overall set of services presented, and ranking services so that the most desirable ones are presented first (automated service composition) [14, 16, 20].

With emerging web-services technology and increased interoperability between multiple distributed architectures, easy to use predefined workflows, survey systems, and federated networks of curate bioinformatics portals were developed. These systems allowed easy access to semantic web service clients for bioinformatics and support systems for service providers [24-29].

Collaborative biological researcher networks and consortia establish comprehensive integrated databases utilising web services architecture to capture genomic information into a comprehensive integrated platform [21]. This facilitated systematic exploration that defined ontologies and generic data models with the relevant communities, to produce standardized nomenclature and data representation. Minimal data models allow simple but broad integration, and inheritance allows detail and depth to be added to more complex data objects without losing integration power [21].

Web services are a key technology for bioinformatics, however incompatibilities between data resources and analysis services exist and some workflows remain non-interoperable. This is not surprising given: 1) the number of emerging life science fields (such as ‘Glycoinformatics’ and interaction networks); 2) the technical challenges involved with large data management and asynchronous services and security. A clear collaborative opening exists to standardize exchange format and services between key stakeholders (web service providers, client software developers and Open Bio* projects) [30]. This cooperation remains critical

between the major database resources and software developers to advance bioinformatics web service technology [30]. However it has been suggested that the primary hindrance to the creation of the SWLS is social rather than technological in nature, dependent primarily on the will and participation of its consumers [15, 19].

2.2.2 Major data resources

Managed under a coordinated effort the three major primary sequence databases exchange data on a peer-to-peer basis under the framework of the International Nucleotide Sequence Database Collaboration [6]. The National Centre of Biotechnology Information, Bethesda, USA (NCBI) exchanges sequence data daily with the European Molecular Biology Laboratory (EMBL/EBI) [31] Nucleotide Sequence Database in Europe and the DNA Data Bank of Japan (DDBJ) to ensure worldwide coverage. NCBI GenBank [32], EBI and DDBJ also maintain and provide numerous database divisions, genome-related viewers, and bioinformatics software (sequence alignment and similarity search tools) [33-35], publications, and Gene expression databases (GEO, SAGE) [9, 36]. All NCBI and EBI databases are accessible through a search application interface NCBI Entrez retrieval system [37], and EBI-eye [38] respectively, these information retrieval systems access data from the major DNA and protein sequence databases along with taxonomy, gene, genome, protein structure, domain information and journal literature. The complete bi-monthly releases and daily updates of the GenBank database are freely available by FTP [32].

GenBank links to outside third party compatible resources, these include pathways [39], Gene Ontology's [23] and major secondary databases such as H-InvDB [40].

Data derived from primary databases can be manually edited to remove errors, duplications and extend annotations to construct secondary databases. The information in secondary databases is therefore considered of a higher data quality. Integrating data from primary and secondary datasets are generally centred based on gene, gene regulation, protein, and disease. The following five points describe data integration resources with centred views of research approaches.

1) Gene centred data integrated sites include, NCBI Entrez Gene, H-InvDB (JBIRC and DDBJ), and Ensembl (EBI and Sanger). Ensembl integrates genomic information for a comprehensive set of chordate genomes. Providing gene annotations for a subset of supported species in addition to specific resources that target genome variation, function and evolution. Ensembl data is accessible in a variety of formats including via a genome browser, API and BioMart [41]. Ensembl also includes assembly, enhanced visualisation and data-mining options for gene regulatory features [11]. EBI BioMart "High-Throughput Data Retrieval" generate annotation tables of large gene sets as input with no statistical analysis for gene ontology predictions [41]. The Biomart web server provides open access through a web interface (MartView), and supports programmatic access through a Perl API as well as RESTful and SOAP oriented web services [41]. This resource integrates functional predictions if available. EBI Biomart has functionality based on the

filtering of genes list via lists of accession numbers, IDs (like Gene division, RefSeq, MIM, InterPro, PDB, GO, Affymetrix), or via their expression (cell types, developmental stages), or via their homology to other species, or via the occurrence of SNPs. The human focused DAVID database for annotation [42], provides visualization and integrated discovery annotation and statistical analysis for GO terms, pathway assignments, and the clustering of orthologous groups (COG, KOG) [43]. PANTHER (Applied Biosystems) classification system was designed to classify proteins (and their genes) in order to facilitate high-throughput analysis for the alternative human Celera genome [44]. Proteins have been classified according to families and subfamilies, molecular functions, biological processes, pathways. The DFCI (TIGR) Gene Indices (TGI) provides species based EST project assembly of tentative consensus sequences and annotation for 41 selected animals, 45 plant species, 15 Protist organisms, and 10 fungal organisms [45].

2) The following are some examples of gene regulation centred resources.
The ENCODE project [46], has a gene regulation centred data integration focus. According to the ENCODE consortium the goal is to identify all functional elements encoded in the DNA sequence. This includes elements that act at the protein level (coding genes) and RNA level (non-coding genes), and regulatory elements that control the cells and circumstances in which a gene is active. Promoters and other transcriptional regulatory sequences, along with determinants of chromosome structure and function, such as origins of replication, remain largely unknown [46,

47]. Regulatory element information is integrated from DNA hypersensitivity assays, DNA methylation experiments [48], and chromatin immunoprecipitation (ChIP-Seq [49]) of proteins that interact with DNA, including modified histones and transcription factors [50]. These results are displayed on the UCSC Genome web based browser for human [8]. The data can also be downloaded to predict potential disease risks, and to stimulate the development of new therapies to prevent and treat diseases.

3) Protein centred research has developed a number of resources dedicated to integrating protein related data. The EBI UniProt Knowledgebase (UniProtKB) is a portal for curated protein information, including function, classification, and cross-references [51]. UniProtKB/Swiss-Prot is manually annotated and reviewed, while the UniProtKB/TrEMBL annotation is automatic and not reviewed. This protein-centred data integration resource, provides access to complete species proteomes; and integrates data from a variety of sources, including UniProt, RefSeq proteins, InterPro, CluSTr, GOA, and ENSMBL Genomes [51].

4) Resources for disease centred research are largely based on human and the model human organisms such as mouse and rat. NCBI ‘s Genetic Association Database (GAD) [52] has human focused disease-centred data integration containing medically relevant polymorphism and mutational data. The standardized nomenclature is dependent on gene nomenclature standardization HUGO [53] and phenotype ontology projects [54]. Genetic association studies, whole genome

association studies (GWAS) and disease genetic models provide information on the genetics of common disease [55].The Online Mendelian Inheritance in Man (OMIM) is the largest knowledgebase of human genes, phenotypes and genetic disorders [56-60]. Tools that will assist researchers to integrate information from resources such as OMIM and NCBI PubMed to identify candidate genes from linkage and association studies are emerging [61, 62].

5) System centred databases and software like other integration resources requires controlled vocabularies for data and information storage, presentation, analysis and modelling of network data (IntAct). Web services allow direct computational access and retrieval of networks in an interoperable format (XML) [63-65]. The Reactome resource provides infrastructure for computation across the biological reaction networks especially for data mining human reactions and pathways [66-71]. System level understanding of protein-protein interactions is available in the STRING database [72-76].

6) Many more specific resources, with a greater diversity of data integrated, are those for an integrated view centred on model organisms. Many secondary database resources exist for the model organisms as a ‘one stop’ access site for genomic and related data, human [40] and mouse [77, 78]. In plant research communities have developed resources to access specific organism information and data, these include the Arabidopsis information Resource (TAIR) [79-84] and Gramene which contain diverse information on the major grass crop species, (rice,

maize, wheat and barley) [85-90]. In invertebrates FlyBase [91-95] and VectorBase are the major resources [96-99], involved in the development of controlled vocabularies for the species of interest [95, 100].

2.2.3 Bioinformatics workflow systems

Sophisticated online research environments are emerging that provide to the end user, easy access and assembly of bioinformatics workflows, and access to Grid and High Performance Cluster Computing [101]. These systems allow the creation and reuse of workflows across distributed computing resources, simplify access to computer and data resources and leave the infrastructure transparent to the end user [101]. Bioinformatics systems available for data integration and mining automate the assembly of bioinformatics workflows. These workflow management systems are required to be extensible and scalable, and be able to access local and remote resources and analysis tools [17, 102-104].

2.3 Genomic analysis and functional predictions

Genomic features associated with diseases at the molecular level can be predicted through data mining and machine learning [105]. Targeting candidate gene features *in silico* decreases the scope of analysis that is required, and improves the accuracy of predictions. This reduces the cost and time involved in analysis [105, 106], by improving model/prediction performance; providing faster and more cost-effective models, and gaining a deeper insight into the underlying processes that generated the data [107].

As Yu in 2008 pointed out, complex diseases can involve the interaction of multiple genes and environmental factors [108]. Data mining using automated workflow

based strategies can significantly improve candidate feature/gene detection and interactions [108]. In concept, a functional genomic analysis, identifies some experimental output within the genome, these outputs are then segmented into defined blocks of initial annotation, followed by clustering into larger derived annotations and networks. Then based on comparative conservation a functional genomic annotation assignment is inferred [109].

In the prediction of protein-ligand interaction sites and the identification of genes in a genomic sequence, a set of manual classified rules are used to create a new set of rules by data-driven machine learning. These new rules are converted into models and applied to classify the statistical properties of unknown cases, Hidden Markov Models (HMMs) [110]. In sequence comparisons, Markov chain amino acid / nucleotide substitution matrices (PAM and BLOSUM) and profile HMMs can be used [111], HMMER is a software example [112]. In a set of homologous sequences with shared sequence motifs new unknown family members can be identified (pFAM) [113]. Gene finding HMM combine many values such as Open Reading Frames (ORFs) and splice regions to identify regions that can be translated and spliced to form a protein (GeneMark.hmm) for specific organisms. HMMs are also used for determining sequence structural predictions, fold assignment and function predictions for cluster sub families [114].

2.3.1 Identification of disease and pathogenic targets

Modelling data using methods, such as HMM, provides reliable quantitative prediction to identify genomic elements involved in disease and pathogenicity

[115]. Statistical methods to associate trait loci with diseases and phenotypes such as quantitative trait loci (QTL) and genome-wide association studies (GWAS) have limited resolution and small heritable variation respectively [116]. Therefore the predictive power and interpretation of QTL and GWAS results consequently have limitations [116] as they are best applied to genetically structured populations and more limited in human population studies. Models with high throughput genomic data can functionally associate genes with phenotypes and diseases to improve the diagnosis and treatment [116].

Immuno-informatics is the application of informatics techniques to the immunity system, for vaccine design, and provides the basis for faster identification of diagnostic targets. This area of study is referred to as “computer aided vaccine design (CAVD) or computational vaccinology” [117, 118]. The prediction of functionally relevant targets based on functional characterisation in biological systems such as host-pathogen interactions and methods for predicting immunogenicity at epitope, subunit or attenuated pathogen levels increases the predictive power for the identification of vaccine candidates [117, 118]. In regards to immuno-informatics the prediction of epitope targets is important. Epitopes are the complementary protein to the Antibodies (Ab) Antigen (Ag) binding sites in Ag-Ab reaction. Epitope sequence can be either conformational (discontinuous) or linear presenting [119]. A number of properties are required for epitope prediction for vaccine design, including that the epitope occur on the surface of the protein and are more flexible than the rest of the protein. The epitope region requires a

high degree of exposure to the solvent and the amino acids making the epitope charged and hydrophilic [120].

In the identification of epitopes many methods have been employed. The Parker Hydrophilic Prediction method [120] utilized 3 parameters, hydrophilicity, Janin's scale accessibility and Karplus & Schultz's flexibility [121]. The hydrophilicity parameter was calculated using HPLC from retention co-efficient of model synthetic peptides. The surface profile was determined by summing the parameters for each residue of a seven-residue segment and assigning the sum to the fourth residue, this is one of the most useful prediction algorithms [120]. Other methods include Emini Surface Accessibility Prediction [122], Chou and Fasman Beta-Turn Prediction [123], exposed surface, polarity and antigenic propensity of polypeptides chains (Kolaskar Tongaonkar Antigenicity) [124] have been correlated with the location of continuous epitopes. Kolaskar semi-empirical method uses physiological properties of amino acid residues and the frequencies of occurrence of amino acids in experimentally known epitopes [124]. B-cell epitopes are recognized by antibodies of the immune system and are used in the design of vaccines and diagnostics tests. It is therefore of interest to develop improved methods for predicting B-cell epitopes [125].

Many immunological databases and analysis resources aid the design and interpretation of information for host pathogen interactions, autoimmune diseases, cancer, transplantation and allergies [126]. In this thesis these data resources were

used to explore, and refine, vaccine targets in ticks. The application of stringent *in silico* criteria and computational workflows allows a more targeted strategy for the identification and prioritisation of potential vaccine candidates lists for testing, reducing both time and cost [127].

2.4 Informatics genomic case studies

2.4.1 Human TE disease

This section reviews current data and tools available for the human transposable element case study. The functional role of human transposable elements and their involvement in disease is then investigated in detail in Chapter 3 (Figure 2.2).

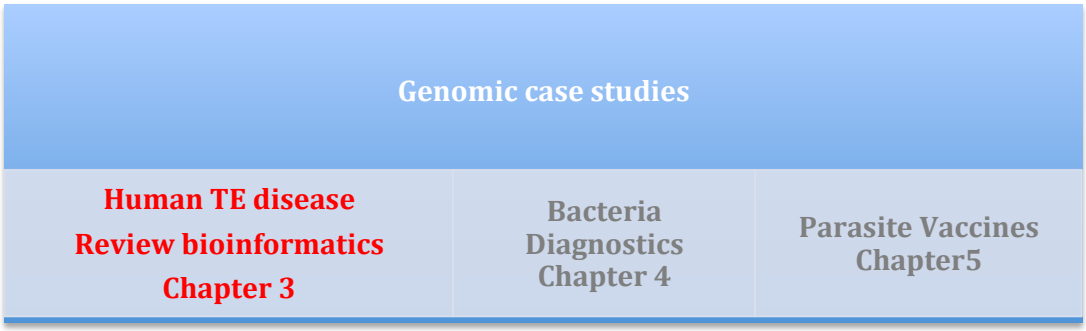


Figure 2.2 Case study one the human TE identification of functional targets in disease

Case study 1: Novel predictions and findings for the possible role of human repeat element in disease.

The *Homo sapiens* Alu elements, a major human repeat have been well researched and families classified. These transposable elements can be identified by sequence alignments with consensus subfamilies datasets as contained in Genetic Information Research Institute (GIRI) RepBase [128] and accessed by

common computer programs such as Repeatmasker3 [129] and Censor [130]. Studies have surveyed the numbers of Alu in human transcript before [131-133], but the overall proportion of Alu containing transcript has remained essentially the same [134].

In case study one, a large amount of heterogeneous sourced transcriptome sequence is available for human. The public H-Invitational (Japan) [135] and NCBI GenBank (USA) [136] databases represent the largest and most complete collections of human transcript sequence information currently available. In H-Invitational secondary database the reference data set for transcript sequence is highly curated with genomic and phenotypic metadata. Sequence tissue and cell type annotation, loci Entrez gene identifier, and the GenBank CDS start and end positions can be found in the JBIRC downloadable files [137]. The NCBI gene identifier and GO term [138] is bridged by data (gene2go.qz) available from the NCBI ftp website [139]. Therefore the human transcriptome and integrated phenotypic data is readily available for sequence analysis to quantify the Alu element content in cancer and determine possible functional roles.
Databases are scalable and SQL compliant to access information from different integrated data types. An open source object-relational PostgreSQL database [140] can be constructed to contain information on (1) H-Inv cDNA such as tissue type and disease trait, the cDNA CDS position, locus Entrez gene identifier and cDNA function (2) cDNA RepeatMasker3 output, (3) NCBI Entrez gene information [141], and (4) GO information extracted from gene ontology version1 with GO slim

text files for the Homo sapiens taxon identifier 9636 [142]. The Alu element can then be investigated within the heterogeneous transcript sequence and metadata to identify Alu content in pathological sourced tissues. The functional role of Alu is presented in Chapter 3 and novel predictions and findings made for the possible role of human Alu repeat element in disease.

2.4.2 Pathogenic Bacteria

This section reviews current data and tools available for the pathogenic bacteria case study. This case study is described in detail in Chapter 4 for the possible functional roles of pathogenic bacteria virulence elements and their use in pathogen diagnostics (Figure 2.3).

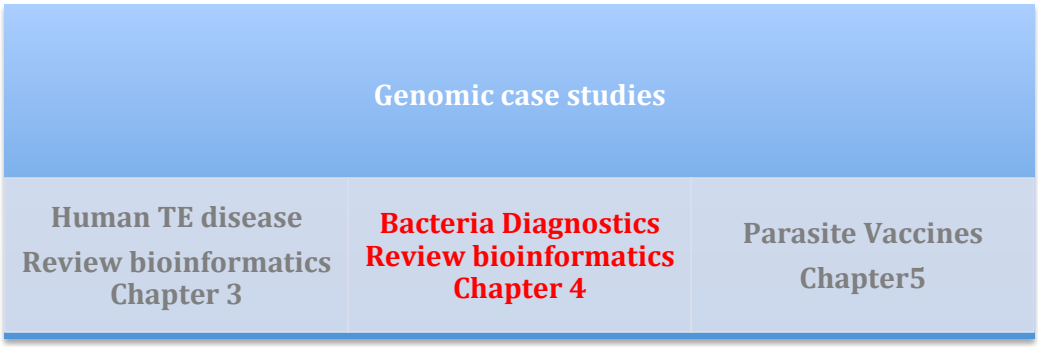


Figure 2.3 Pathogenic bacteria case study to identify functional targets in disease

In case study 2, pathogen genomes can be comparatively analysed for the identification of gene targets for *Campylobacter* diagnostics. In the second case study, a *Campylobacter fetus* subspecies *venerealis* (*Cfv*) that can cause venereal disease in cattle is compared to a close sub species reference genome to identify strain specific targets for diagnostics. Comparative analyses between the two

subspecies are conducted for the identification of gene targets for *Campylobacter fetus* diagnostics.

The *Campylobacter fetus venerealis* genome (2Mb) is partially available. A subset of 273 *Cfv* contig sequences (lengths greater than 2Kb) from 1,187 of the assembled contigs was available from the UNSAM, Argentina for this analysis. The assembled contigs have been submitted to GenBank as a part of the WGS division (GenBank: ACLG00000000 and RefSeq: NZ_ACLG00000000). Fourteen *Campylobacter* genomic reference sequences are available for comparative analysis from NCBI RefSeq Genome (http://www.ncbi.nlm.nih.gov). The sub species *Campylobacter fetus* subspecies f*etus* 82-40 is the closest reference genome available to *Campylobacter fetus* subspecies *venerealis*. All genomic detail for other *Campylobacter* species is available from NCBI (http://www.ncbi.nlm.nih.gov/genomes) genome division. Comparative genomic assembly of *C. fetus* subspecies *venerealis* based on the reference genome of *Campylobacter fetus* subspecies f*etus* 82-40 is possible through sequence similarity programs such as BLAT [143]. Comparative viewers such as Argo [144] enable quick genomic / gene overviews to identify the genomic regions shared and those unique to each genome.

Open sourced tools such as Glimmer3 are available to predict open reading frames (ORF) and/or genes [145]. The annotation of predicted ORF sequences can be estimated from a similarity search with the BLAST program [146] against public

NCBI protein (nr, patent), String [147], COG [43], and NCBI Conserved Domain databases [148]. The results from Blast searches can then be readily parsed and categorised using BIOPERL [149] scripts based on alignment percent identity (PID) and target/query percent coverage thresholds. The functional categories of proteins can be determined based on the String Database [147] categories that have been developed upon the NCBI COG database role descriptions. The main categories being Cellular processes and signaling, Information storage and processing, Metabolism, Poorly characterized, No mapping, Non Orthologous Group (NOG) and KOG (euKaryote Orthologous Group). Many genes that play important roles in pathogenesis and host specificity is relatively unknown, for bacteria such as *Campylobacter* many gene functions remain unknown or hypothetical [150].

2.4.3 Ectoparasites and vector borne disease control

This section reviews current data and tools available for the parasite vaccine case study. The functional genomic target identification in a large ectoparasite (tick) for vector control is described in further detail in Chapter 5 (Figure 2.4).

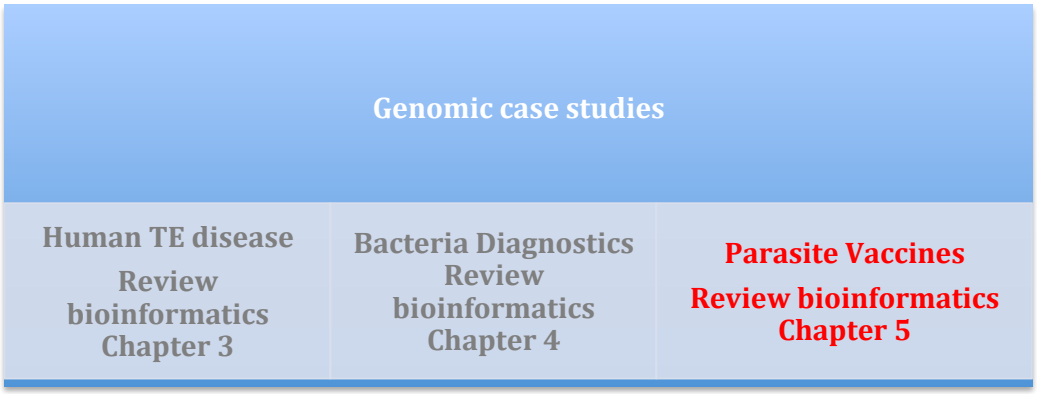


Figure 2.4.Ectoparasite case study to identify functional targets in vector control

In case study 3 *de novo* gene target predictions are made in a large complex genome. The hard tick *Rhipicephalus microplus* is a serious pest and vector of disease for cattle. The cattle tick's large, complex non-sequenced genome (7.1Gb) has available transcript data sets from a number of libraries [151, 152]. Only one hard tick genome, *Ixodes scapularis,* has been sequenced and is available for comparative analysis [153]. This genome is too distant to aid in the assembly of *Rhipicephalus microplus* but is valuable for the identification of gene sets. Through comparative analyses and data integration, genomic targets can be identified. Specialised arthropod resources that contain sequence and phenotypic data include VectorBase (*Ixodes*) [96-99] and FlyBase (*Drosophila*) [91-95, 100, 154-161].

The choice of genome assembly tools to use depends on the sequence technology used. Sanger sequence (800bp read) assembly applications include phred/phrap [162], CAP3 [163], Phusion [164] and MIRA [165]. For the Pyrosequencing (454) Genome Sequencer (GS), the Newbler application assembles 300bp reads [166], Genome Analyzer (GA) sequencing by synthesis (Solexa) short reads use de Bruin graph assemblers such as ABySS [167], Velvet [168] and Euler-SR [169]. ABySS is the only tool to use message-passing interface (MPI)-cluster approach.

An arthropod repeat data set (library) is available through Censor [170] and RepeatMasker3 [171]. However, the repeat sequence of under characterised / *de novo* assembled genomes in contrast to the human case study (see section 2.4.1)

require *de novo* repeat identification [172]. Once the repetitive elements are identified and masked the gene content can be predicted using eukaryote open source applications such as GlimmerHMM [173] and Genscan [174] and predictions trained for new models. Annotation can then be conducted as described in case study 2 (section 2.4.2) for the predicted protein and transcript sequence. Further resources are available to identify candidate epitope peptides (Table 2.1). These tools are reviewed for high throughput (HTP) analysis accessibility.

Site	Prediction	URL	HTP
ABCPred [175]	B-Cell	http://www.imtech.res.in/raghava/abcpred/	No
Antigenic	B-Cell	http://emboss.sourceforge.net/apps/release/5.0/emboss/apps/antigenic.html	Yes
BCEPred	Linear B-Cell	http://www.imtech.res.in/raghava/bcepred/index.html	No
BCIPep	B-Cell blast database	http://www.imtech.res.in/raghava/bcipep/data.html	Yes
Bepipred	Predict location of linear B-Cell epitopes	http://www.cbs.dtu.dk/services/BepiPred/	Yes
CEP [176]	Conformational Epitope Prediction Server B-Cell	http://bioinfo.ernet.in/cep.htm (no longer available)	No
IEDB	B-cell and Tcell helper epitope tools	http://tools.immuneepitope.org/main/html/bcell_tools.html	No

Table 2.1 Summary of B-Cell epitope prediction tools reviewed

3 Chapter three - The transcript repeat element: the Human Alu sequence as a component of gene networks influencing cancer

The contents of this Chapter have been published as a review [177].

3.1 Introduction

Transposable elements (TEs) contribute to about half the content of the human genome and have been found in abundance in gene sequences and in a significant portion of mature mRNAs [178]. Most of the TEs in the human genome are retroelements, such as the LTR, Alu, L1 and MIR sequences that have spread throughout the genome by transcription, retrotranscription, retrotransposition and insertion into various locations [178]. Retrotransposition and recombination of TEs can contribute to human molecular evolution, generating mutations associated with inherited disease, defining the organization of the genome into active and inactive regions, suppressing transcriptional noise and regulating transcript stability [179].

3.1.1 The functional roles of human TE elements

TEs can influence gene expression by a number of methods previously reviewed [180-182]. Feschotte described the ways TEs can influence gene expression at the transcriptional and posttranscriptional levels. At the transcription level an inserted TE can, provide an alternate gene promoter (transcription start site), disrupt or introduce new gene cis-regulatory element(s), drive antisense transcription, and act as a site for heterochromatin silencing [180].

At the posttranscriptional level a TE inserted into the ‘three prime’ (3′), untranslated region (UTR) can introduce, an alternate polyadenylation site, and provide a target site for microRNA (miRNA) or RNA-binding protein. Pre-mRNA intron TEs can produce alternative splicing (intron retention, exon skipping) [180]. Alu elements inserted in opposite orientation have been reported to undergo base-pairing and affect the splicing patterns of downstream exon, shifting it from constitutive to alternative coding [183]. Pre-mRNA intron TEs can also be incorporated (‘exonized’) as an alternative exon. This results in the translation of a new protein isoform or in the destabilization or degradation of the mRNA via the Nonsense Mediated Decay (NMD) pathway, especially if the exonized TE introduces a premature stop codon [180].

It is possible that genomic TE families can recruit the same regulatory motif(s) at many chromosomal locations, creating a multiple gene regulatory network [180, 184-187]. It is also suggested that chromodomains could be responsible for the targeted integration of LTR retrotransposons, favourable for mobile elements, and avoiding negative selection from insertion into coding regions [188].

TE families can be linked into a small RNA network, when co-transcription of the TE-host gene with a small RNA precursor containing a TE of the same family fold into a near perfect palindromic structure (as seen for MITEs) forming double-stranded RNA. This is then processed into a mature miRNA, and TE-derived

miRNA can then pair with complementary TE sequences embedded within the 3′ UTR of co-transcribed mRNAs [180].

Small non-protein-coding RNAs (ncRNAs) are important components in the regulation of eukaryotic gene expression [189]. Several classes of small regulatory RNA, including miRNAs, small interfering RNAs (siRNA), repeat-associated small interfering RNAs (rasiRNAs) and piwi-interacting RNAs (piRNAs), can use partially overlapping pathways similar to RNA interference (RNAi) to silence gene expression, by degradation or translation inhibition of mRNAs containing complementary sites [180]. A fully integrated small RNA pathway connecting ncRNA entities such as miRNAs, siRNAs, trans-acting small interfering RNA (ta-siRNAs), and a natural antisense transcripts (NATs) pathway involving naturally occurring siRNAs (nat-siRNAs) have been reconstructed within *Arabidopsis thaliana* [181]. The relationship of piRNAs, siRNAs and rasiRNAs to TEs has been reported [181] and the proposed function of these small RNA is to silence invasive DNA from viruses and control the replication of TEs [190, 191]. Several mammalian miRNA precursors have been found to contain or be derived from TE sequences [192, 193] and a substantial number of predicted miRNA targets map within members of the same TE families [192-194], again pointing at a model whereby large sets of cis-regulatory sequences have been seeded by transposition [180]. Approximately 12% experimentally characterized human miRNA genes have originated from TEs [193]. Transcriptional gene silencing of elongation factor 1 alpha (EF1A) has been shown in human tissue culture cells to inhibit mRNA

transcription by promoter-directed siRNA [195]. Small RNA gene silencing can occur through mRNA cleavage, translational repression, and transcriptional repression through the modification of DNA and/or histone, and DNA elimination through the modification of histone [182].

Post-transcriptional regulation of gene networks by a single small RNA shared cis-element is similar to transcriptional regulation by transcription factors [180, 196], this is supported by small RNAs homology-dependent transcriptional silencing and their participation in the nucleation of heterochromatin [180, 191, 197]. Small RNAs are also involved in the phenomenon of paramutation a homology-sensing mechanism of interaction described in several kingdoms, including human. It was hypothesised that pairing interactions between specific chromatin complexes and trans-RNA based communication are two non-mutually exclusive models [198, 199]. A series of tandem repeat sequence upstream of the b1 transcription site activate the biosynthesis of flavonoid pigments in plant, when highly methylated at the *B-I* allele was dark purple relative to *B'* light purple, an inheritable change in phenotype [200]. A RNA-mediated trans-induction of chromatin mechanism was proposed for the role for small interfering RNA (siRNA) [201]. The dsRNA that is formed by transcription from the two strands of the repeated DNA is a target for Dicer, which produces siRNA. The siRNA is then believed to mediate chromatin changes, which alters the expression of the adjacent gene. These mechanisms include, but are not limited to, RNA-directed DNA methylation and RNA-directed histone modification [198].

The dsRNA generated from the repeats by sense and rare antisense transcription, produce siRNAs that could be efficiently amplified from tandem array transcripts by RNA dependent RNA Polymerase (RdRP) and Dicer. The RdRP activity, and complementary strand RNA using siRNA primers synthesis, results in increased amounts of siRNAs throughout the repeats. The production of dsRNA that trigger RNAi, can then result in: degradation of homologous mRNA; altered chromatin states associated with DNA methylation; or potentially, inhibition of translation [198].

3.1.2 Disease targets in human TE elements

As described earlier almost half of the human genome is composed of transposable elements derived from exogenous genetic invaders, most related to retroviruses [202]. Two main classes are DNA transposons and the retrotransposons. Retrotansposons are comprised of three groups, the autonomous, long terminal repeat (LTR) ERV and non-LTR L1, and nonautonomous SINE and SVA.

Due to the replication mechanism of retroelements hundred of thousands to millions of near identical DNA copies make favourable ground for rearrangements making retrotransposons a strong evolutionary force. A small number of LINE1s (L1s) are active, and move their own and SINE (nonautonomous elements that do not encode protein) sequences into new genomic locations and occasionally causes disease. There are 65 known human disease causing insertions of L1, Alu

and SVA [203].

Retrotransposon recombination can cause deletions, duplications, or rearrangements of gene sequence, Alu elements alone have been implicated in almost 50 disease causing recombination events [203]. Endogenous reverse transcriptase activity has been detected in some tumors and patients with certain pathologies.

It has been previously discussed that changes in global methylation, and the consequent pattern changes of the transcriptome, frequently involve hypomethylation of Alus, LINEs, and HERVs, however demethylation leading to increased endogenous retrotransposition was untested [179, 204]. Furthermore although untested it was possible that elevated retrotransposition within a developing cancer could alter the tumor phenotype or hasten tumor progression [179, 204]. An L1 insertion into a tumor suppressor gene has been reported in a case of colon cancer [205]. The synergy between retrotransposition and cancer is an unexplored question; and that by profiling *de novo* insertion events in tumor versus normal cells valuable information on mobile element regulation would be gained [179, 203].

Among the various retrotransposable element families, the primate specific Alu, so called because it contains a recognition site for the restriction endonuclease AluI [206], is the largest family of short interspersed nucleotide elements (SINEs). Alu

has an estimated 1.3 million copies contributing to 10% of the human genome [133, 178].

Alu elements are dimeric sequences with a characteristic length of 300 base pairs that probably originated from a gene encoding the 7SL RNA [207] and then developed into dimeric sequences via the Alu monomeric forms FAM, FRAM and FLAM [208, 209]. The full-length 300 base pair (bp) Alu element has evolved in primates into a number of different dimeric families and subfamilies that are distinguished from each by diagnostic markers within their sequences as well as by their evolutionary history or phylogeny [210]. Essentially, the three main Alu families or classes categorized as Alu-J, Alu-S and Alu-Y were estimated by sequence divergence to have evolved specifically in primates about 65-80 million years ago (mya), 30-50 mya and 15-30 mya, respectively [210]. Each of the three Alu families has additional subfamily members that can be identified by sequence alignments with consensus subfamilies using the computer program Repeatmasker3 [129], Censor [130] or other programs [133]. The Alu family found most commonly in transcripts and the human genome are from the AluS subfamilies [133]. Studies have surveyed the numbers of transcript [131, 132], but the overall proportion of Alu containing transcript has remained essentially the same [134].

Fragmented and/or full-length Alu elements have been found in the coding regions of mRNA and may be beneficial, neutral or deleterious to the function of the gene

and its transcripts. Alu RNA levels have been reported to increase in response to cell stress [211, 212]. Alu sequences are known to regulate gene expression and translation at the transcriptional and posttranscriptional levels [213], modulate cellular growth, differentiation and tumour suppression [214] and function in exonization [215]. For example, Alu elements are a source of adenine and uracil rich elements at the 3' UTR that may contribute to the stabilization or degradation of mRNA [216]. Approximately 5% of alternately spliced internal exons in the human genome were found to have an Alu sequence [215], with most Alu exons alternately spliced and with only a segment of the Alu sequence contributing to the new open reading frame (ORF) [217].

Earlier studies suggested that Alu transcription is regulated by epigenetic mechanisms such as DNA methylation and histone modification at Alu repeats [218, 219]. It has been found in gastric cancer that chromatin remodeling at Alu element repeats by DNA demethylation and histone deacetylase inhibition (HDI) can activate expression of Alu-associated miRNAs, which can down regulate target oncogenes in human [220].

3.2 Material and Methods

3.2.1 Databases and resources

In order to better understand the TE content of transcripts and their possible functions in human, two of the public transcript databases the H-Invitational (Japan) [135] and NCBI (USA) [136] human transcript datasets were analyzed. These two data sets are the largest and most complete collections of human transcript sequence information currently available. The JBIRC Version 3.6 H-InvDB from the H-Invitational human full-length cDNA annotation project (dated September 28th 2006) provided 167,992 cDNA accessions representing 35,005 cDNA clusters or 'transcribed genes' or loci.

3.2.2 Sequence ontology

All sequence annotation on tissue and cell type, loci Entrez gene identifier, and the GenBank CDS start and end positions were extracted from the NCBI GenBank [136] using Bioperl tools [221] and JBIRC downloadable text file FCFUN [135]. The H-InvDB functional annotation categories are also conveniently available as follows.

Category I: identical to a known human protein with greater than or equal to 98% identity and 100% coverage.

Category II: similar to known protein in another species with greater than or equal to 50% identity.

Category III: containing an InterPro [40] protein domain. Category IV: conserved hypothetical protein, greater than or equal to 50% identity to a known hypothetical protein.

Category V: hypothetical protein with an ORF length greater than or equal to 80 amino acids, no pseudogene overlaps.

Category VI: Hypothetical short protein, with an ORF length less than 80 amino acids.

Category VII: pseudogene candidates.

3.2.3 Bioworkflows

Bioperl [221] and Perl [222] scripts were written to parse tissue and disease information from the H-Inv text files and the GenBank records.

The tissue types and traits (diseased or normal) were defined and categorized using Perl scripts with a defined list of text string synonyms for different tissue and cell traits and types to extract the origin of the tissue or cell. Tissues were categorised by organ type for this study.

3.2.4 Alu Analysis Within the cDNA

All cDNA sequences were screened for Alu repeat elements using the computer program RepeatMasker3 [129]. The RepeatMasker3 outputs provided the Alu positions within the cDNA, Alu size, Alu subtypes and the Alu percentage sequence similarity to the consensus sequence.

3.2.5 Gene Ontology

To bridge the gene identifier and GO ontology terms [138] the text file gene2go.qz (18/10/2006) was downloaded from the NCBI ftp website.

3.2.6 PostgreSQL DB

To query the different data types an object-relational PostgreSQL database was constructed. The database cDNA-Alu contains tables with information on (1) H-Inv cDNA such as tissue type and disease trait, the cDNA CDS position, locus Entrez gene identifier and cDNA function (2) cDNA RepeatMasker3 output, (3) NCBI Entrez gene information, and (4) GO ontology information extracted from gene ontology version1 with GO slim text files for the Homo sapiens taxon identifier 9636. More than 115 H-Inv 3.6 cDNA Entrez gene identifiers were updated due to being recently discontinued or replaced in NCBI. The database was queried to find all cDNA containing Alu and their tissue source and disease trait state (Appendix 3.1). The PostgreSQL database tables can be found in Appendix 3.2.

3.2.7 Mapping H-Inv Loci

The chromosome positions for cDNA loci were extracted for all the trait states and then mapped to the Human Genome assembly version 36.1 in the CMap [223] application to view genomic patterns in loci between cDNA normal and disease trait.

3.3 Results and discussion

3.3.1 Bioinformatics workflow

The bioinformatics approach to determine the Alu element content in human transcript involved eight steps (Figure 3.1). The first and second filters removed all sequences derived from cell culture, then those sequences not sourced or annotated from an organ respectively. The third filter classified sequences sourced from disease and the fourth filter retained only those annotated sequences sourced from cancer. In step five the data sets are statistically compared to analyse the Alu content, locations and function in the transcript data sets, steps six and seven respectively. In Figure 3.1 those sequences filtered to black boxes were eliminated from further analysis. The red box highlights the proposed Alu feedback model.

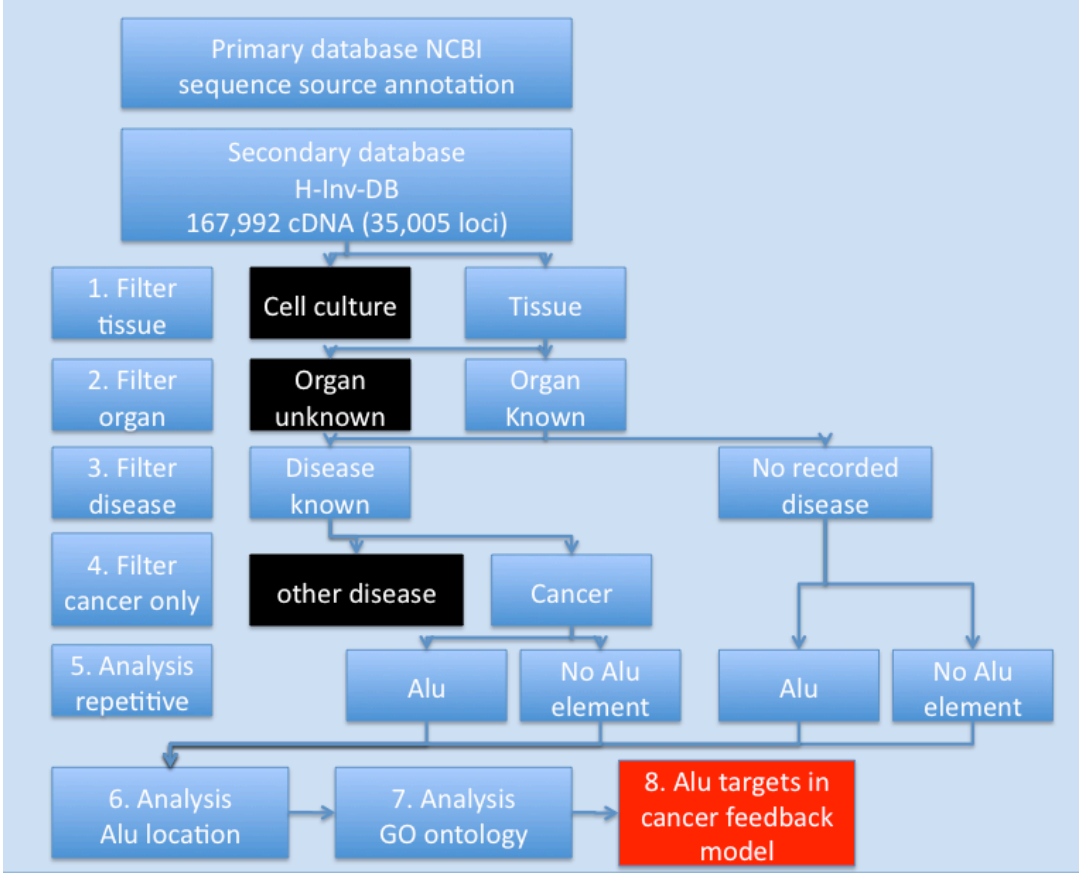


Figure 3.1 Bioinformatics workflow approach for the Alu element analysis in human transcriptome

The bioinformatics workflow analysis steps are discussed in further detail in section 3.3.2.

3.3.2 Differentiating between the cancerous and normal tissue information within the integrated H-Inv database

The H-Invitational human full-length transcript annotation project provides transcript accessions representing 35,000+ loci in the H-Invitational Database (H-InvDB). This integrated data has been highly curated by experts to provide high-

quality manual and computational annotation of human genes and transcripts and information about gene structures, alternative splicing isoforms and non-coding functional RNAs [224].

To ensure that the H-Invitational and NCBI human transcript datasets were of sufficient quality for investigation, general filters were applied as follows:
(1) Transcripts derived from culture cells were removed and only those found in tissues retained.
(2) Association of transcripts with cancerous tissues was determined by evaluating a match to the following 22 terms: adenocarcinoma, astrocytoma, carcinoma, choriocarcinoma, glioma, glioblastoma, gliosarcoma, hepatoma, leukaemia, lymphoma, melanoma, melanotic, neuroblastoma, neuroepithelioma, papilloma, pheochromocytoma, retinoblastoma, rhabdomyosarcoma, teratocarcinoma, tumour.
(3) Tissue information was then categorised by organ type.
(4) The dataset sizes for normal or cancerous tissues were examined and under represented tissues excluded from further analysis.
(5) The transcript length qualities were evaluated in order to assess the Alu positional insertion within the transcript.
Transcripts can be divided into three regions, a five prime untranslated region (5' UTR), coding sequence (CDS) and a three prime untranslated region (3' UTR), the Alu content represented based on quantified base pair percentage and the fraction of total transcript numbers. To further assess the quality of the transcripts between

the two categories of normal and cancerous tissues, the 3'UTR regions in housekeeping genes were investigated as EST sequencing was biased against this region. The average length of the 3'UTR sequence for the 149 expressed genes was 558.3 nucleotides for transcripts derived from cancerous tissues and 957.9 nucleotides for transcripts derived from normal tissues, and the t-test significant ($p<0.05$) at $p = 0.045$. Since the statistical difference between the normal and cancerous tissues for the length of their 3'UTR was relatively borderline, it was assumed that the transcripts were of sufficient quality in both cancerous and normal tissues to undertake the following further comparisons.

3.3.3 Human Alu repeat sequence as a component of gene networks

The development and public access to primary and secondary databases based on sequence information gives an opportunity to investigate the distribution of Alu in transcripts. Previous analyses of Alu-transcript have resulted in the creation of databases that highlight Alu positions within transcript [131].

In order to better understand the possible regulatory functions of Alu within RNA, surveys of full-length transcript from a number of normal and cancerous tissues, were used to compile transcripts encompassing all the nucleotide sequences from the CAP site to the poly (A) addition site or at least the entire coding sequence of a protein. The H-Invitational database (H-InvDB) provides 167,992 full-length and partial transcripts encoded by 35,005 human gene loci [225]. The loci gene transcript structures and alternative splicing forms, referred to here as isoforms can be found at H-InvDB website [135].

For the total 167,992 H-Inv transcript sequences, 107,001 sequences are derived from a tissue, those allocated to a gene loci and the number of Alu containing transcript (Alu-transcript) in normal and cancerous tissues were then calculated. Table 3.1 shows the relative numbers and percentages for the distribution of the gene or genomic loci containing transcripts with and without Alu sequences. It was also determined if these loci transcripts were either unique or common to both the cancerous and normal tissues. For the total 106,825 transcripts, 33,918 transcripts expressed from 13,798 loci were assigned to a cancerous tissue trait and 72,907 transcript expressed from 26,677 loci were assigned to a normal tissue trait. The number of Alu-containing transcript (Alu-transcript) totaled 17,861 (17%) of the 106,825 transcripts derived from tissues and was represented by 13,240 loci or 44% of the total gene loci. This set of Alu-transcript is represented by 14,191 normal-sourced tissue sequences (10,648 loci) and 3,670 cancer-sourced tissue sequences (2,592 loci), represented by 19% and 11% of normal and cancerous sourced tissue sequences, respectively. For the represented 29,979 loci, 3,302 loci (11%) produce transcript derived only from cancerous tissues, 16,181 loci (54%) produced transcript only from normal tissues and 10,496 (35%) of loci contained transcript produced from both cancerous and normal tissues.

Table 3.1 Percentage and number of unique and overlapping loci groups for loci containing transcript and Alu-transcript sequences in normal and cancerous tissues.

	# Loci	Normal only	Cancer Only	Both Normal and Cancer	# Transcript	# Alu-transcript	%	# Loci with Alu-transcript	% Loci with Alu-transcript
Normal condition	26,677	16,181	-	10,496	72,907	14,191	19	10,648	39.91
Cancer condition	13,798	-	3,302		33,918	3,670	11	2,592	18.79
Total		29,979			106,825	17,861	17	13,240	44.16

The fraction and relative tissue distribution of Alu-transcript to transcript within the cancerous or the normal tissue group is shown in Figure 3.2. Of the 25 tissues examined, 22 had a statistically significant proportional difference of Alu-transcript to transcript in the normal to the cancerous tissues as determined by Fisher's exact two sided tests ($P <= 0.001$ for 14 tissues; and $P <= 0.01$ for 8 tissues), Appendix 3.3. The four cancerous tissues with a significantly ($P < 0.001$) greater proportion of Alu-transcript to transcript in them than in the normal tissues were liver, oral cavity, ovary, and placenta. Of these four tissues, the greatest relative difference between the normal and cancerous tissue for Alu-transcript to transcript was in the ovary. Esophagus, pancreas and skin showed no significant ($P > 0.05$) difference between normal and cancerous tissues. Of the cancerous tissues, the greatest proportion of Alu-transcript to transcript was in the oral tissues, whereas, of the

normal tissues, the greatest proportion of Alu-transcript to transcript was in the rectal tissues.

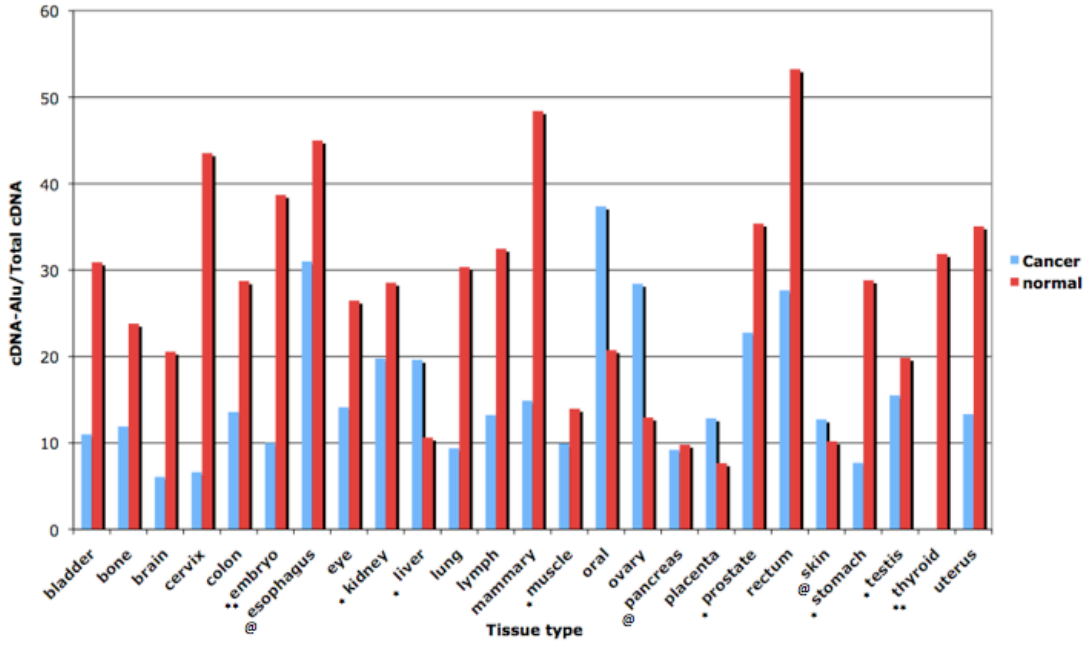


Figure 3.2 Frequency of Alu containing transcript (Alu-transcript) in human normal and cancer tissue types. The frequency of Alu-transcript per transcript sequences represented in normal (blue bars) or cancerous tissues (red bars). The statistical significance of the differences between the normal and cancerous tissues by a Fisher's exact two sided test is at a probability (P) less than or equal to 0.001. Symbols indicated next to the tissue names are * P <=0.005, ** P <= 0 .01, @ = Not significant.

The tissue results are consistent with previous studies that reported that the RNA embedded with Alu showed variable patterns between tissue types, but that there was an overall two fold increase of edited Alu-transcript in normal cells relative to malignant cells [226]. Significant hypo-editing of the Alu elements within the transcripts of tumors from the brain, prostate, lung, kidney and the testis has been

reported and it was suggested that A-to-I RNA editing was an epigenetic mechanism relevant to cancer development and progression [227]. In their tissue analysis, they found that the placental tissue was an exception in that there was more editing of Alu within the transcripts of cancerous than normal tissue. Recently based on the results obtained for 9 different editing sites, it was determined that RNA editing is an epigenetic mechanism that does not participate in the evolution of urinary bladder cancer [228]. A significantly higher proportion of Alu-transcript to transcript in normal tissues than in the corresponding cancerous tissues of the bladder may suggest another mechanism for bladder cancers. Further, there is a significantly higher proportion of Alu-transcript to transcript in cancerous tissues than in the corresponding normal tissues of the liver, oral cavity, ovary and placenta. This may suggest that the Alu-transcript oncogenic affect may occur in most, but not all tissues and as reported may not be completely reliant on A-I editing in certain cases.

3.3.4 Alu-transcript functions

The functional categories of the Alu-transcript collected from the H-Inv human transcript database were analysed by Gene Ontology (GO), which classifies the known functions of gene products into at least 5,175 categories according to biological processes, cellular components and molecular functions [138]. GO analysis was performed for transcripts with a NCBI gene identifier and CDS to determine the different metabolic and regulatory pathways that are possibly associated with the Alu-transcript in normal and cancerous tissues. Over 8,000 Alu containing transcripts (Alu-transcripts) from the H-Inv database were classified into

nearly 3000 of the GO functional categories, and then assigned into 35 GO Slim categories for the Homo sapiens taxa identifier 9636. Over 70% of the Alu-transcript loci grouped into binding and catalytic activity molecular functions, and more than 80% were involved in biological processes, such as physiological process, nucleotide metabolism, cell communication and transport (Appendix Figure 3.4). Less than 3% of the transcript and Alu-transcript loci were designated "molecular function unknown" or "biological process unknown". The molecular function, biological process and cellular component of loci unique to tissue traits for the full transcript and Alu-transcript subsets showed no Alu-transcript affect on the GO slims [229] category proportions. The proportion of loci unique to tissue traits for a few molecular functions did however vary between unique normal and cancer tissues for all transcripts and Alu-transcript. The molecular functions of both transcript and Alu-transcript were associated mainly with binding (DNA) (40%) and catalytic activity (20%). The number of loci products involved in the function of nucleic acid binding increased nearly two-fold for the disease loci, whereas loci associations with signal transducer activity and transporter activity decreased in the cancerous trait three and two fold, respectively. These shifts in function were significant by a Fisher's exact test at $P < 0.0001$.

3.3.5 Alu families and subfamilies

When Alu elements were categorized into their particular families and subfamilies, the contribution of the Alu families in transcripts was similar to the proportion of Alu families within the genome; AluS 54%, AluJ 26%, AluY 10%, monomeric 8% and

Alu 2% [133]. The most frequent Alu subfamily was AluSx. No family bias was detected between the cancerous and normal tissue sets (Appendix 3.1).

3.3.6 Alu locations within the transcript

The relative position and size of Alu insertions within the transcript sequences was determined for 17,819 of 17,861 distinct transcript sequences with an annotated CDS that contained at least one Alu. The Alu positions were partitioned into three groups according to the regions in which they were found (1) overlapping the coding sequence (CDS), (2) exclusively in the 5' UTR, or (3) exclusively in the 3' UTR. The number of fragments and distinct transcripts in normal or cancerous tissues, and transcripts with known functions are indicated in red in Appendix Figure 3.5. The transcript Alu content was measured based on the proportion of transcripts containing an Alu fragment represented within the CDS, 5'UTR or 3'UTR or represented by sequence length (base pairs) to the total sum of sequence represented in these three regions (Table 3.2).

Table 3.2 Transcript with Alu number, percentage and lengths (bp) represented in available transcript 5'UTR, CDS and 3'UTR regions.

	Number sequences								
	Total	**Normal**	**Cancer**	**Alu-transcript**	**%**	**Normal Alu-transcript**	**%**	**Cancer Alu-transcript**	**%**
5' UTR	99671	67868	31803	6585	6.61	5619	8.28	966	3.04
CDS	106415	72645	33770	3930	3.69	3203	4.41	727	2.15
3' UTR	101292	69163	32129	14228	14.05	11224	16.23	3004	9.35

	Seq. length (bp)								
	Total	**Normal**	**Cancer**	**Alu-transcript**	**%**	**Normal Alu-transcript**	**%**	**Cancer Alu-transcript**	**%**
5' UTR	35897377	29061960	6835417	1716495	4.78	1526474	5.25	190021	2.78
CDS	103506598	70328008	33178590	507946	0.49	426050	0.61	81896	0.25
3' UTR	82313000	61388458	20924542	4250911	5.16	3445289	5.61	805622	3.85

3.3.7 Alu sequence quantification within transcripts

Greater than 28,000 Alu fragmented or full-length elements were found in the 17,000 Alu-transcript sequences surveyed, at an average of 1.6 Alu per sequence. The Alu elements ranged in size from10 nucleotides in length to full-sized. All these Alu fragments were within the RepeatMasker [171] recommended RM score to be outside the threshold for a false positive match. Appendix Figure 3.6A shows the number of Alu fragments plotted as a percentage of the consensus Alu sequence length for transcripts from cancerous and normal tissues, with known and unknown (hypothetical) functions. The plots show that the number of Alu-transcript in normal tissues is clearly higher than in cancerous tissues. In the Alu-transcript plot, minor peaks represent a slightly increased number of the monomeric form of the Alu structure (110-130 bp) and major peaks above 85% of the Alu consensus full-length sequence represent a greatly increased number of the dimeric form of the Alu structure (>250 bp). This trend of minor and major peaks was similar for the Alu-transcript in both the normal and diseased tissues. Appendix Figure 3.6B shows the number of Alu fragments plotted as a percentage of the consensus Alu sequence length for transcripts from the different Alu families, AluS repeats are highly represented by lengths greater than 85%.

While multiple fragments of Alu elements can overlap the CDS within a single transcript, the largest number of Alu contained in the CDS was 4 copies in one

particular transcript sequence. Full-length dimeric Alu (>280 bp) fragments that overlap the CDS, that represent half of the transcripts, potentially have cryptic splicing sites provided by the Alu sequences.

In an examined subset, the majority of the transcript isoforms containing full-length Alu within the coding regions of the transcripts were categorised as hypothetical proteins rather than known proteins. As the function of these Alu-transcripts is not known or is labelled ‘hypothetical’, it is proposed that they are either untranslated or destined for rapid degradation or they serve a binding role against excessive retrotranspositional activity within the genome (Appendix 3.7).

3.3.8 Incorporation of Alu elements as part of transcribed genes

The size of the Alu within transcripts is surprisingly mostly full-length (Appendix Figure 3.5) without open reading frame disruption. The high prevalence of expressed full-length Alu, greater than 85% of the full element, in more than half available loci supports the view that the Alu within transcripts might have some important function rather than random insertions within transcripts with no molecular or biological function [211, 227, 230]. One important role for the complementary Alu transcript sequences may be to participate in RNA sense/antisense hybridisations [212] or dilution effects to protect the genome from unbridled Alu retrotransposition events by inhibiting or limiting the expression of the active Alu master copies involved with Alu retrotransposition. Overall, Alu within transcripts contribute significantly to various characteristics and patterns in transcriptome activity. Some examples of genes that have transcripts with Alu

sequences embedded in the 3'UTR are glucose -6-phosphatase, placental alkaline phosphatase and receptors for platelet activating factor, vitamin D, interferon-alpha, epidermal growth factor and various other cytokines and interleukins. Consistently a larger proportion of Alu containing transcript is found in those mRNA derived from normal than cancerous tissues regardless of the Alu sequence position within the mRNA (Table 3.2). Some of the genes with Alu in their 5' UTR are METTL8, LIPT1, BDKRB1, GBA, KCNH5 and SLC39A1. The orientation of Alu fragment within the transcript is predominately antisense in the 5' UTR and coding sequence (CDS) regions and sense in the 3'UTR.

In an earlier study of 87 Alu containing transcripts, only 4 Alu were found to overlap the CDS [134]. We found in our review of the H-Inv transcript database that 3,639 of the 17,861 Alu-transcript sequences contained at least one Alu fragment that overlapped with the CDS. The distribution of the Alu within the CDS was determined as any overlap in the 5' end of the CDS or the 3' end of the CDS, spanning the entire CDS or contained entirely within the CDS (internal). From these distributions over 200 transcripts contained an Alu that spanned the entire CDS.

3.3.9 The impact of Alu elements within the transcript UTR

The relative proportion of Alu that overlapped the CDS in our study was increased approximately 10 fold above the numbers previously provided [134], and over half of the Alu sequences overlapping the CDS were greater than 85% of full-length elements. A significantly higher number of Alu sequences were present in the 5'

and 3' UTR regions than the CDS (Table 3.2). Therefore the Alu location within the transcript is strongly biased to the 5' and 3'UTR end of the transcript (Table 3.2). This finding is in contrast to an Expressed Sequence Tag (EST) study on cell lines, where 82% of transposable elements that are found within the EST derived from cancerous tissues were located in the CDS [230]. The Alu sequences within the 3' UTR may affect mRNA stability or degradation by A to I editing in Alu sequences [231] or by contributing adenine and uracil rich elements (ARE) to the transcript [216]. This Alu-associated RNA editing is a potential mechanism for marking non-standard transcripts for degradation rather than for translation [232]. The Alu sequences within the 3' UTR might also affect translational efficiency by providing secondary and tertiary structures to hinder translational editing or translational rates [212, 216]. The Alu sequences within the 5' UTR of transcripts provide cryptic promoter sites, steroid binding sites or other regulatory elements and also secondary or tertiary structures that may hinder or enhance translation of the transcript [212]. The BRCA1 Alu-rich gene is an anti-oncogene involved with a hereditary predisposition to ovarian and breast cancer is a well-studied example of a gene that expresses variable forms of a transcript with and without an Alu insertion in its 5'UTR sequence [233]. Some of the genes with Alu in their 5' UTR sequences, such as METTL8, LIPT1, BDKRB1, GBA, KCNH5 and SLC39A1, however, might exploit the same Alu regulatory mechanisms as BRCA1 in disease [233].

3.3.10 The impact of Alu elements within transcribed exons

The Alu sequences within genes can contribute to different transcriptional isoforms by providing intron-exon recognition sites, exonization [234-236]. Approximately 5% of alternately spliced internal exons in the human genome were found to originate from an Alu sequence [215]. Most Alu derived exons are alternately spliced and only a segment of the Alu contributes to the new ORF [237]. In this review we found that 16% of the Alu-transcript transcripts provided potential splice sites within CDS. Coding sequence from some functioning human genes exist that are almost entirely derived from Alu elements, such as the AD7C gene that encodes a neuronal thread protein and has 99% of its transcript composed of 4 to 5 Alu fragment elements [238, 239] although the validity of these findings has been questioned on the basis of EST and genomic sequence analysis [240]. It was further argued that functional proteins are unlikely to contain transposable cassettes derived from young transposable elements, but if so, then their role is probably limited to regulatory functions [241]. Some of the genes previously found to have Alu sequences contributing to exonization include ADARB1, DSCRB8, ITCH, CDK5RAP1 [132], RPE2-1, C-rel-2, MTO1-3 and PKP2b-4 [217]. Potentially, numerous new cryptic splice sites exist in the human transcriptome [242] and many of them await full structural and functional characterization.

The CDS base pair coverage by Alu is nearly 40% in over a half of the Alu-transcript with an Alu size greater than 280 base pairs. Of the 3,639 transcripts, 2,473 were annotated as hypothetical proteins. To determine if the Alu content

changed the location or structure of the 'normal' CDS, a subset of transcript with full-length Alu sequences internal to the CDS were further investigated. In this subset 85 percent of the Alu-transcript were annotated as translating hypothetical products with no known function. The remainder were either identical to the known genes glucose-inhibited division protein A family protein and Penicillin-binding protein, dimerisation domain containing protein or similar to the genes, N4BP2, Alpha-COP, A1BG, dNT-1, GPI transamidase component PIG-U, RPIP8, PAOX, programmed cell death 6, enkurin, CRL2, G protein-coupled receptor 43, selenoprotein N precursor, CTAGE family member 5 isoform and a solute carrier family 24 (sodium/potassium/calcium exchanger), member 5 (Appendix 3.1, Appendix 3.7).

The variation in the A1BG gene transcripts provided an example of how the Alu content of the transcripts might change their structure, function and tissue specificity of expression. The protein encoded by the eight exons of the A1BG gene is a plasma alpha-1glycoprotein with sequence similarity to the variable regions of some immunoglobulin supergene family member proteins. A1BG interacts in the plasma with the cysteine-rich secretory protein 3 (CRISP-3) that is secreted by neutrophilic granulocytes and it is believed to play a role in innate immunity. Some variants of the A1BG transcripts were transformed by the Alu insertions from full-length coding forms (1645 bp – 1810 bp) composed of eight exons into longer transcripts (1951 bp – 3466 bp), usually with shorter ORF of 1 to 3 exons in the coding region. The Alu containing A1BG transcripts were found in

the amygdala, cerebellum and tetracarcinoma, while the Alu foetal and adult livers, primary hepatoblastoma and ovary expressed Alu free A1BG transcripts.

3.3.11 Alu-siRNA mediated feedback model in disease

The transcriptome-wide survey of human transcript, carried out in this Chapter represented nearly 30,000 gene loci, which is close to the full complement of known gene loci of the human genome. Overall, 13,240 loci (44%) expressed a transcript that contains a partial or full-length Alu sequence. Alu fragments are far more abundant in the transcriptome than previously reported. No discernable bias for Alu families was identified between the normal and cancerous tissues, which confirmed that the Alu families in the transcriptome reflect the proportion of Alu families in the genome. For all the total transcript analysed, 17% contain an Alu fragment which is four times greater than previous estimates of Alu-transcript content made from smaller data sets and based on different search methods [134].

Loci-based transcriptome-wide investigation of Alu sequence has highlighted that Alu fragments are proportionally more abundant in normal tissue transcript than the corresponding cancerous tissues, many with hypothetical functions. In our feed back model (Figure 3.3) we propose that a high proportion of the non-functional (hypothetical) Alu-transcript expressed in normal tissues or during cellular stress and infection does not undergo degradation, whereas the Alu-transcripts are degraded in cancerous tissues due to increased Alu small RNA activity.

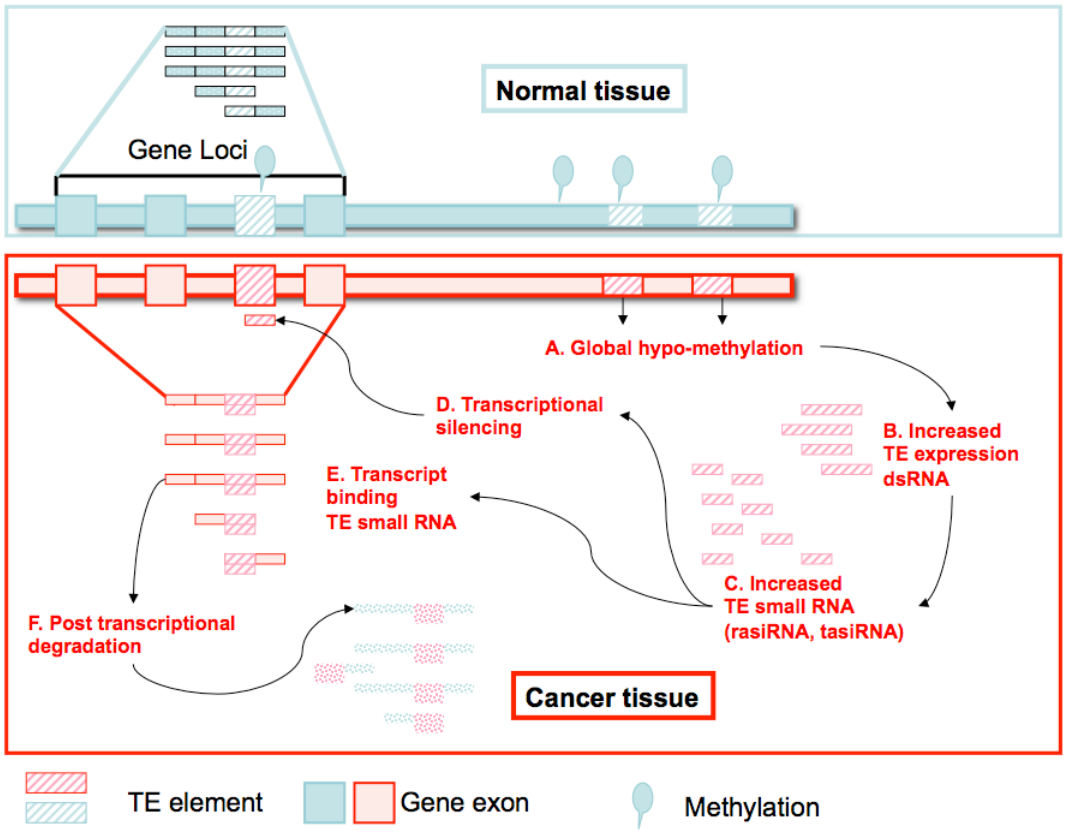


Figure 3.3 Feedback model for Alu mediated siRNA interference of Alu containing transcripts in cancerous tissues

Loss of genome-wide methylation is a common feature of cancer. Bollati in 2009 hypothesized that DNA methylation initially evolved as a defence mechanism against viral and other DNA pathogens as a way to silence foreign DNA sequences [243-245] This is consistent with the observation that LINE and SINE (Alu) transposable elements, are heavily methylated in normal cells [246]. It has been found that global hypomethylation of LINE-1, Alu and SAT-alpha is significantly associated with tumour progression in *Mus musculus* and may contribute toward a more extensive stratification of the disease [247]. Our model proposes that in cancer tissues the hypomethylation of intergenic Alu-loci may result in transcription

of these elements to generate Alu-dsRNA molecules that are further cleaved into siRNAs. The generated Alu-siRNAs guide the RNA Silencing Complex (RISC) to Alu-containing mRNAs and mediate their posttranscriptional degradation (Figure 3.3).

Methylation and chromatin structure together play a role between retroelements and their host. Hypomethylation and expression in developing germ cells opens a "window of opportunity" for retrotransposition and recombination that contribute to human evolution, but also inherited diseases. In somatic cells, the presence of retroelements may be exploited to organize the genome into active and inactive regions, to separate domains and functional regions within one chromatin domain, to suppress transcription noise, and to regulate transcript stability. Retroelements, particularly Alu, may fulfil physiological and protective roles during responses to stress and infections [179].

In support of our model it has been previously reported that reduction of the methylation index of L1 and Alu following treatment of three lung cell lines with 5-aza-2'-deoxycitidine, consistently resulted in increased expression of both elements. This study demonstrated a strong link between hypomethylation of transposable elements with genomic instability in non-small cell lung cancer and provided early evidence for a potential active role of these elements in lung neoplasia. As demethylating agents are now entering lung cancer trials, it was viewed imperative to gain a greater insight into the potential reactivation of silent

retrotransposons in order to advance the clinical utilization of epigenetics in cancer therapy [248].

To investigate the role of microRNAs (miRNAs) on epigenetic therapy of gastric cancer, the miRNA expression profile was analysed in human gastric cancer cells treated with 5-aza-20-deoxycytidine (5-Aza-CdR) and 4-phenylbutyric acid (PBA). Microarray miRNA analysis shows that most of miRNAs activated by 5-Aza-CdR and PBA in gastric cancer cells are located at Alu repeats on chromosome 19 [220]. Analyses of chromatin modification showed that DNA demethylation and HDAC inhibition at Alu repeats activates silenced miR-512-5p by RNA polymerase II. These results suggest that chromatin remodeling at Alu repeats plays critical roles in the regulation of miRNA expression and that epigenetic activation of silenced Alu-associated miRNAs could be a novel therapeutic approach for gastric cancer [220].

Hypomethylation of the genome largely affects the intergenic and intronic regions of the DNA, particularly repeat sequences such as transposable elements, and believed to result in chromosomal instability and increased mutation events [249]. It has been considered that global demethylation of repeat sequences including transposable elements and the site-specific hypomethylation of certain genes might contribute to the deleterious effects that ultimately result in the initiation and progression of cancer and other diseases [249]. In our model the increased hypomethylation of transposable elements is shown in the lower section of Figure 2

highlighted in red to represent the cancerous tissue state, while normal state is represented in the upper half of the diagram in blue.

In the feedback model genomic hypomethylation (Figure 3.3) and the increase in TE sequences is shown correlated with the increased production of small RNA. The small RNA complementation to loci containing TE may then silence transcription and the production of mRNA (rasiRNA) and/or post-transcriptional binding TE containing transcripts (tasiRNA) leading to their degradation, mRNA cleavage. Mature miRNA sequences of approximately 50 additional human miRNAs have been shown to lie within Alu and other known repetitive elements, extending the current view of miRNA origins and the transcriptional machinery driving their expression [250]. Further, base-pair complementation can be demonstrated between the seed sequence of a subset of human miRNAs and Alu repeats that are integrated parallel (sense) in mRNAs [251].

3.4 Conclusion

The key findings from this chapter are that TEs such as Alu elements can influence gene regulation/expression on both the transcriptional and post-transcriptional level (Figure 3.3). TEs are found to be an abundant source for small RNA and hence potential gene interference activity. In examining the available human transcript data in H-Inv, the Alu element was found with a much higher abundance in transcript than previously reported, serving as possible gene targets for repeat derived small RNA (Alu-siRNAs). Due to the curated nature of H-Inv, transcripts could be assigned to their tissue source, disease condition and function.

Transcripts from cancerous tissues revealed an underrepresentation of transcript containing Alu elements, and the majority of full length Alu containing transcripts derived from normal tissues were highly represented by hypothetical proteins. It was proposed that Alu derived small RNA interference activity increases in certain cancerous tissue with increased genome hypomethylation, thus suppressing the abundance of Alu derived transcripts. The proportion of Alu-transcript was significantly higher in 18 of the 22 normal tissue types than in the corresponding cancerous tissues (Figure 3.1). Therefore the Alu-transcript oncogenic effect may occur in most, but not all tissues and as reported may not be completely reliant on A-I editing in certain cases. The hybridised complementation between the small RNA and the repeat containing transcripts in hypomethylated diseased tissue may lead to reduced RNA-Alu transcript levels due to their degradation and/or interference (Figure 3.3).

In summary the key findings from the developed bioinformatics approach for this chapter are: 1) Bioinformatics approaches have provided access to sequence and metadata for novel analysis; 2) through the integration and mining of sequence metadata, sequence was charted to organ and cancer for repeat element and functional analysis; 3) An over representation non cancerous Alu transcript with unknown function was found; 4) Alu transcripts were under represented in cancer sourced transcripts in the majority of tissues; 5) A model for an Alu transcript mechanism in cancer was presented.

4 Chapter Four - Genomic analysis of *Campylobacter fetus* subspecies: identification of candidate virulence determinants and diagnostic assay targets

The contents of this chapter have been published [252]. The laboratory-wet chemistry that is reported in this chapter was carried out by cited co-authors in order to test the bioinformatics outputs from the work.

4.1 Introduction

Bacterial species level genome diversity has been attributed to mobile genetic elements (MGE) and bacteriophages, that confer new characteristics to the recipient genome [253]. The flexible gene pool of bacteria includes mobile and accessory genetic elements, such as bacteriophages, plasmids, pathogenicity islands, insertion sequence (IS) elements, transposons and integrons [254]. MGE such as plasmids (extra-chromosomal self-replicating DNA molecules) exist in cells as replicons and can insert into the chromosome [254].

IS element DNA sequences, are generally less than 2.5 Kbp in length, that encode functions involved in their translocation, and transposition both within and between genomes. IS elements are a subset of a general group of elements named transposable elements. Transposons DNA molecules code for a transposase (and other genes), and are flanked by inverted repeat DNA sequences. Conjugative transposons also carry genes related to plasmid-encoded conjugation, are able to transfer between cells via conjugation [254]. Bacteriophages (Prokaryote-infecting

viruses), can modify the host coding genome, and are capable of inserting into the genome (prophages) [254]. The final genetic elements Integrons, encode an integrase gene and cassette-associated genes [255-257].

Large bacterial chromosomal regions referred to as genomic islands (GI) contain clusters of functionally related genes or operons that are flanked by direct repeat sequences, these are located near an integrase/transposase gene and a tRNA gene [253, 254].

Pathogenic IS genes, are known to be involved with antigenic variation of surface exposed proteins, and environmental adaptation [258, 259]. These IS transfer and integrate pathogenic islands (PAI) into the genome [259-261]. Common insertion sites for PAI are tRNA genes [254].

4.1.1 Bacterial virulence factors

Pathogenomics is the genomic study of pathogenic microbes and how they interact with a host. Pathogenic microbes sense, infect and then colonize their hosts [262]. Microbial pathogenesis/virulence is the measure of the likelihood of a pathogen to cause infection, therefore important for the control of disease [263]. Virulence factors (VF), the genetic elements (gene products) required by the pathogen to colonise, proliferate and cause damage in a host have many known mechanisms [264]. These known virulence mechanisms include motility, chemotaxis, adhesion, invasion and toxin production and regulation by two-component systems [150]. Other VFs such as secreted protein (protein toxins and enzymes) and cell-surface

structures (capsular polysaccharides, lipopolysaccharides and outer membrane proteins), also contribute to the microbes disease process in the host [264].

Virulent secretion mechanisms such as, siderophores (high-affinity iron chelating compounds), catalases, and regulators are indirectly involved in pathogenesis, and are important for the bacterium to establish infection [264]. The Type III secretion system, for example, injects virulence factors into the host cell [265].

4.1.2 Virulence targets for diagnostics

The development of vaccines is recently based on a composition of prominent immunogenic parts of microorganisms (subunit vaccines) or genes. Traditional vaccines incorporate whole live attenuated or killed microorganisms. These have limited application due to concerns about safety, efficacy and/or ease of production, particularly for use in humans [253].

It is therefore possible from a genome sequence, using known prediction properties, to select potential peptide sequences for the construction of chemically synthesised vaccines based on peptide or DNA. The peptide approach can be used directly or used to construct a composite protein made from individual epitopes [253].

Common genomic targets for vaccines include those genes encoding outer membrane proteins or lipoproteins, transmembrane domains or export signal peptides. These surface-exposed or secreted proteins as well as virulence factors

(such as toxins or adhesive factors) can induce an immune response that may be protective for the host [266, 267].

Common target type epitopes are B cell, helper T lymphocyte and cytotoxic [253]. T lymphocyte and B cell targets have been made, and improved methods are constantly being developed [253]. These recombinant proteins from identified ORFs can be produced and screened for distribution in different serotypes, stability, immunogenicity and cross-protection tests [268-271].

Northern Australian beef herds have a 35% unexplained reduction in calf production. In Argentina, calf production has not declined, but remains at a constantly low rate (63%). Causative agents include *Campylobacter fetus* subspecies *fetus* and *venerealis*. To aid the detection and treatment of cattle infected with *Campylobacter fetus* our genomic analysis has identified candidate subspecies specific genes that can be used as diagnostic tools.

The *Campylobacter* genus is a Gram-negative, spiral-shaped bacterium and includes 23-recorded species in the NCBI Taxonomy division. *Campylobacter* spp. colonise diverse hosts from livestock to humans with varying degrees of virulence [150]. Hosts include cattle, swine, bird, and can be the major cause of human bacterial gastroenteritis [272]. *C. fetus* subsp. *venerealis* (*Cfv*) is the causative agent of bovine genital campylobacteriosis, which causes conception failure and embryo loss, with bulls acting as asymptomatic carriers [273]. *C. fetus* subsp. *fetus*

(*Cff*) causes infertility and infectious abortions in domesticated sheep, goats and cattle [272]. It is also an opportunistic pathogen in humans that can severely affect immuno-compromised patients. Initially the bacterium can cause gastroenteritis, and then spread systemically throughout the blood (bacteremia) and cause septicaemia, meningitis, and other systemic infections [272]. Bovine genital campylobacteriosis is an Office International des Epizooties (OIE) notifiable disease considered to have socio-economic and public health implications, particularly with respect to the international trade of animals and animal products [274] .

Although *Campylobacter* sub species have largely conserved genomes, sub species display variable virulence phenotypes in animal models and this phenotypic virulence has been speculated to be due to hyper-variable antigenic diversity and immune evasion [150, 275]. Very few gene targets have been identified for the differentiation of *C. fetus* subspecies, with members of the subspecies shown to be 86% similar based on Pulsed Field Gel Electrophoresis PFGE-DNA profiles [276]. Diagnostic testing of *C. fetus* colonies from transport medium and the biochemical differentiation of the 2 subspecies *venerealis* and *fetus* is important for the diagnosis of bovine venereal disease in cattle. *Cff* and *Cfv* can be differentiated from each other using a range of biochemical assays including H_2S, selenite reduction, growth at 42^{o}C, susceptibility to metronidazole and cefoperazone, basic fuchsin, $KMnO_4$ and glycine tolerance [276, 277]. Glycine tolerance is the OIE recommended assay. It is however difficult to isolate viable

colonies from transport medium for biochemical analysis due to prolonged transport, contaminant overgrowth and the fastidious nature of the bacteria [278-280]. In addition doubts in regard to the stability of these biochemical markers has been suggested based on evidence from phage transduction [276, 281-283]. The H_2S test although described as differentiating *Cff* (positive) and *Cfv* (negative), a *Cfv* strain subsequently named *Cfv* biovar *intermedius* is positive in this assay [284]. Molecular typing methods such as amplified fragment length polymorphism (AFLP) and multilocus sequence typing have been developed to differentiate *C. fetus* isolates [281, 285], but these methods require the isolation of pure colonies which are impractical for diagnostic application. Specific polymerase chain reaction (PCR) assays have been designed and applied to detect *Cfv* [286-288], however it has been suggested that the gene targets are plasmid borne and that in some cases have not reliably detected all *Cfv* isolates [289].

A sensitive real time assay designed to target the *parA* gene originally targeted by PCR assay [287], identified a high prevalence of *Cfv* in Australia cattle not associated with venereal cases [274]. It was thus postulated that isolates of *Cfv* differ in virulence and that other methods may be required to confirm the presence of pathogenic *Cfv* in clinical samples. Genomic *Campylobacter* comparisons of *C. fetus* subspecies and a list of *Cfv* specific genes will provide the basis for developing specific diagnostic assays and improving our understanding of *C. fetus* virulence and epidemiology. No studies to date have reported the putative identification or extensive analysis of *Cfv* virulence genes.

Based on comparative analysis on recently available genome data for both *C. fetus* subsp. *venerealis* (*Cfv*) (incomplete) and *C. fetus* subsp. *fetus* (*Cff*) the work in this chapter has developed a number of assays targeting virulence factors previously identified in *C. jejuni*, *C. coli*, *C. lari,* and *C. upsaliensis* genomes. These virulence mechanisms include motility, chemotaxis, adhesion, invasion and toxin production and regulation by two-component systems, as discussed by Fouts et al. in 2005 [150]. This paper provides the first detailed analysis of available genome sequences in order to identify targets for differentiating *C. fetus* subspecies. Based on the analysis, several targets were identified and confirmed using PCR assays.

Our aims were to (1) identify and compare *C. fetus* putative virulence genes, and (2) characterise genomic features to differentiate the highly conserved *C. fetus* subspecies for diagnostic assays. The genomic features of *Campylobacter* provided subspecies markers that discriminate *C. fetus* species and subspecies, in particular the *C. fetus* sub species (*Cfv* and *Cff*) from each other and other *Campylobacter* species.

4.2 Materials and Methods

4.2.1 Bacterial strains, culture conditions and DNA preparation

Campylobacter fetus subsp. venerealis AZUL-94, an Argentinean field strain isolated from a bovine aborted fetus in 1994, was grown routinely on Tryptic Soy Agar plates or in Brain Heart Infusion (BHI) and cultivated under microaerobic

conditions in anaerobic jars with CampyGen envelopes (OXOID) at 37°C. Total DNA from *Campylobacter fetus venerealis* was isolated by the classical SDS/proteinase K/Phenol/Chloroform extraction method [290]. The Pfizer strains were originally isolated by CSIRO Australia [291].

4.2.2 Library construction, DNA sequencing and assembly

Genomic DNA was randomly sheared by nebulization, treated with Bal31 nuclease and blunt ended with T4 DNA polymerase. Fragments were size fractionated by agarose gel electrophoresis and ligated to dephosphorylated *Hinc*II-digested pBS plasmid. Three libraries with insert size of approximately 2 Kbp (Cf1), 4 Kbp (Cf2), and 6 Kbp (Cf3) were generated. Template preparation and DNA sequencing were performed as described [292] from randomly selected clones. Single-pass sequencing was performed on each template using T7 or T3 primer. Sequencing reads, obtained from the three genomic libraries (Cf1, Cf2, Cf3) were masked against plasmid vector and base called with phred (-trim_qual). Those sequences with at least 50 good quality bases after trimming were retained for assembly. After reaching ~ 4.5X shotgun coverage, assembly was done using the phredPhrap script provided with phrap. The autofinish functionality of consed was used to select candidate clones for re-sequencing to increase sequence coverage, decrease the number of contigs and increase the consensus quality in a number of cases. Additional information on *Campylobacter fetus venerealis* sequencing can be found in Appendix file 4.1.

The genomic sequence data have been deposited in the WGS division of GenBank under the following accession numbers: ACLG01000001…ACLG0101187

4.2.3 Genomic data

A subset of 273 *Cfv* contig sequences (lengths greater than 2Kb) from 1,187 the assembled contigs was generously supplied by the UNSAM, Argentina for this analysis. The assembled contigs have been submitted to GenBank as a part of the WGS division (GenBank: ACLG00000000 and RefSeq: NZ_ACLG00000000). All manuscript referenced contig ORFs are listed in the Appendix 4.2 and 4.3.

Completed *Campylobacter* genomic sequences were obtained from NCBI RefSeq Genome [293]. All genomic sequences for the genus *Campylobacter* listed were downloaded from NCBI genome division, 28th April 2008. *Campylobacter* species included *C concisus* 13826, *C. curvus* 525.92, *C. fetus* subsp. *fetus* 82-40, *C. hominis* ATCC BAA-381, *C. jejuni* RM1221, *C. jejuni* subsp. *doylei* 269.97, *C. jejuni* subsp. *jejuni* 81-176 and *C. jejuni* subsp. *jejuni* 81116.

Alignment of *Campylobacter* genomes was conducted using BLAT [143]. The BLAT results were then filtered for a minimum 50% alignment coverage at greater than 90% identity. The two *C. fetus* subspecies were then displayed in Argo [144].

4.2.4 Alignment of genomic *Cfv* contigs based on *Cff*

The 273 *Cfv* AZUL-94 contigs were aligned to the *Cff 82-40* genome (NC_008599) using BLAT [143] (>90% identity). *Cfv* contigs were assembled (order and orientation) based on the best BLAT alignments between *Cfv* and *Cff* into a contiguous pseudomolecule. Unaligned contigs were concatenated to the pseudomolecule linear sequence.

4.2.5 *Cfv* Open reading frame identification & annotation

ORF prediction was conducted on the 273 *Cfv* contigs using Glimmer3 [145] for ORF lengths greater than 100 nucleotide bases, resulting in 1474 open reading frames (ORF), Appendix 4.4. The 273 *Cfv* contigs and 1474 ORF were subsequently screened against NCBI protein (nr, patent), String [147], COG [43], and NCBI Conserved Domain databases with the BLAST program [146]. These results were then categorised using BIOPERL [149] scripts based on alignment percent identity (PID) and query coverage to provide the following six alignment categories, (1) known protein > 80% PID and > 80% query coverage, (2) known protein > 30% PID and > 80% query coverage, (3) hypothetical protein > 80% PID and > 80% query coverage (4) hypothetical protein > 30% PID and > 80% query coverage, (5) alignments with an expected value less than 1e-05, < 30% PID and < 80% query coverage, and (6) alignments greater than 1e-05 < 30% PID, < 80% query coverage.

4.2.6 *Campylobacter* protein similarity to *Cfv* ORF

Campylobacter complete proteome sequence and protein detail were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). The 8 complete campylobacter proteome sets were compared to our *Cfv* ORF set using BlastMatrix [294] at an ARL 0.75 and an expected value less than 1e-05 (results in Appendix 4.5).

4.2.7 Putative virulence genes

The functional categories for *Cfv* ORFs were determined based on the String Database [147] categories developed on NCBI COG database role descriptions. The main categories being Cellular processes and signaling, Information storage and processing, Metabolism, Poorly characterized, No mapping, Non Orthologous Group (NOG) and KOG (euKaryote Orthologous Group). The ORFs identified in *Cfv* were screened against the String database and alignment results were filtered using Bioperl for greater than 80% query coverage and 30% PID or with an expected value less than 1e-05. These ORFs were then screened against NCBI protein database to determine selected putative virulence gene representation in the *Campylobacter* genus.

4.2.8 Primer design

Primer sets were designed using EMBOSS [295] Primer3 [296] on *Cfv* putative virulence genes and genes unique to *Cfv*. Primers were screened against the Cfv AZUL-94 strain and *Cff* (strain 82-40) genome data and public databases to confirm specificity. Assays were conducted in 20 μl reaction volumes, using 10nM of each forward and reverse primer, 1 x PCR reaction buffer with 25mM Mg^{2+} (HotMaster *Taq* buffer, Eppendorf, Germany), 200μM dNTPs, 1U Hotmaster™ *Taq* DNA polymerase and 1 ng of *C. fetus* DNA. The reactions were cycled in a Gradient Palm Cycler (Corbett Research, Australia), using the following temperature profile: an initial denaturation at 94^{o}C for 2 min, followed by 35 cycles of denaturation at 94^{o}C for 20s, annealing at 45 to 57^{o}C (dependent on primer pair, Table 4.1) for 10 s, and extension at 72^{o}C for 30s including a final single extension

for 7 min at the end of the profile. Amplification products were separated in 2% TBE (89 mM Tris borate, 2 mM EDTA, pH 8) agarose gels using 100bp ladder (Invitrogen) and were visualised under UV illumination by ethidium bromide staining. DNA preparations from strains were screened in all assays.

4.3 Results

4.3.1 Bioinformatics workflow

In Figure 4.1 the assembled genomic contigs of *Cfv* were scaffolded into a Phase Two (ordered and oriented contigs) genomic assembly (pseuochromosome/molecule) using the complete genome of *Cff*. The *Cfv* gene predictions were then compared to the *Cff* protein and NCBI NR datasets and functional ontology assigned. The unique *Cfv* gene set as compared to *Cff* were then presented as candidate diagnostic targets for PCR analysis.

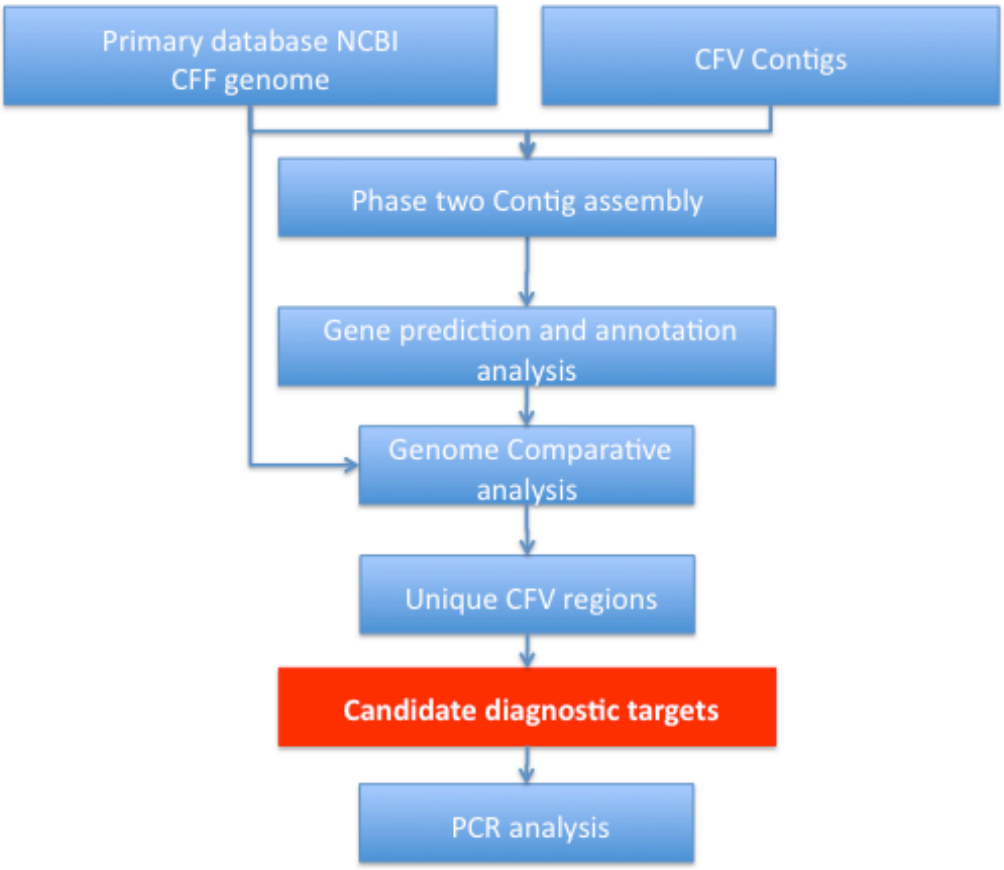


Figure 4.1 Bioinformatics workflow for the identification of candidate diagnostic genes for discriminating campylobacter subspecies

The bioinformatics workflow analyses are described in further detail in the following sections.

4.3.2 Assembly of *Cfv* for identifying targets for diagnostics

The available genomic sequence information (ca 75-80% *Cfv* genome) was compiled using the complete *Cff* 82-40 genome sequence (NC_008599) in order to identify diagnostic targets for the detection of *Cfv*. The ordering of available

genome segments generally aligned well with the *Cff* genome as shown in Figure 4.2

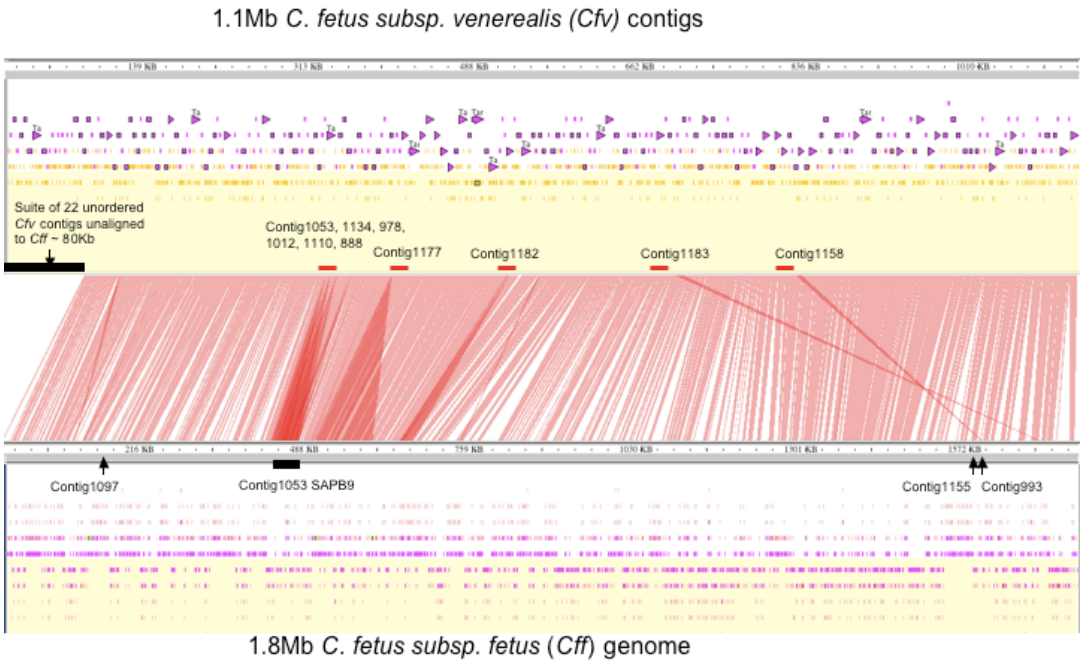


Figure 4.2 Genomic nucleotide alignments of *C. fetus* subsp. *venerealis* (*Cfv*) contigs to the *C. fetus* subsp. *fetus* genome. Genomic nucleotide comparison of *C. fetus* subsp. *venerealis* (*Cfv*) contigs (1.08Mb) as aligned to the *C. fetus* subsp. *fetus* (*Cff*) completed genome (1.8Mb). Orange shaded regions between the parallel sequences of *Cfv* (top) and *Cff* (bottom) highlight contigs in common; unshaded regions are unique between the two *Campylobacter* subspecies.

Several striking features were evident in the subspecies comparison. Firstly, an 80 Kb suite of 22 *Cfv* specific contigs (relative to *Cff*) housed a range of putative virulence factors such as Type IV secretion systems (Appendix 4.2). Secondly a number of potential virulence factors were also identified in the genomic sequences that were shared between *Cfv* and *Cff* (Appendix 4.3). Table 4.1 summarises virulence factors by comparing the ORFs of the 2 *C. fetus* subspecies with 4 *Campylobacter* species as described previously [150]. In general similar

numbers of genes potentially associated with 2 component systems, toxin production, outer membrane proteins, and motility were identified. Only one bacterial adherence gene was identified in both *C. fetus* subspecies with 2 and 3 ORFs identified in *Cfv* and *Cff* respectively (Table 4.1). However an additional adhesion homologue fibronectin (FN) binding ORF was identified in *Cfv* but not *Cff*. A large surface array protein was found highly conserved between the two subspecies as evident in the genomic sequence alignments (Figure 4.2).

Table 4.1 *C. fetus* subsp. *fetus* (*Cff*) and subsp. *venerealis* (*Cfv*) virulence factors compared with 4 other *Campylobacter* spp.

Putative virulence type	**Other spp.**[a]	***Cff***	***Cfv****
Bacterial adherence	9	3[b]	4[b]
Motility	55-66	41	46
Two-component system genes	11-15	16	14
Toxin and resistance	15-20	9[c]	7[c]
Membrane proteins	185-218	209	202

Summary of *C. fetus* virulence gene ORFs in *C. fetus* subsp. *fetus* (*Cff*) and subsp. *venerealis* (*Cfv*) compared with 4 other *Campylobacter* spp. (adapted from Fouts et al).
[a] *C. jejuni, C. lari, C. upsaliensis, C. coli* (Fouts *et al.* 2005)
[b] *Cff* – PEB1 (3) – no other adherence homologues found; *Cfv* ORFs – PEB1(2), *cadF*(0), *jlpA* (1-poor homology), Fibronectin binding (1), 43-kDA MOMP (0)
[c] not including resistance genes for *Cff* and *Cfv*, toxin subunit ORFs only
*N.B. *Cfv* genome incomplete

The nucleotide alignment of *Cfv* contigs to the closest sequenced genome *Cff* displayed those *Cfv* contig sequence in common between the two genomes (not specific to *Cfv*) and *Cfv* contig sequence not found in *Cff* (specific to *Cfv)* (Figure

4.2). Of the 273 *Cfv* contigs, 251 contigs (993569 bp) were conserved with *Cff* and 22 contigs (86999 bp) specific to the *Cfv* genome compared to *Cff*. Contigs specific to *Cfv* were Contig1018, Contig1021, Contig1023, Contig1024, Contig1030, Contig1031, Contig1042, Contig1120, Contig1139, Contig1165, Contig1181, Contig1185, Contig1186, Contig419, Contig733, Contig846, Contig851, Contig872, Contig875, Contig914, Contig958 and Contig991 (note all ORF without strong homology to *Cff* are listed in Appendix 4.2).

When probed against all available genome protein sequence information the *Cfv* specific contigs (Table 4.2) had the following alignments; two contigs (~4.9 Kb) with short alignments to only non-campylobacter bacterial species (Contigs914 and 875) (*Campylobacter* specific); five contigs (~20 Kb) with significant alignments to *C. jejuni* and *C. coli* plasmid genomes and short alignments to *C. hominis* and *C. lari*; ten contigs completely unique to *Cfv* (*Cfv* specific) (~32 Kb); and five contigs (~27 Kb) with significant protein alignments to *Cff* although this was not evident at the nucleotide sequence level.

Table 4.2. Comparison of *C. fetus* subsp. *venerealis* specific contigs (length bp below) to selected available genomic protein sequences. Alignment information reported as 'excepted value; score and percentage identity' below

Campylobacter fetus venerealis	Bacillus licheniformis ATCC 14580, complete genome	Helicobacter pylori plasmid pHel5	Campylobacter jejuni subsp. jejuni 81-176 plasmid pTet	Campylobacter jejuni strain 81-176 plasmid pTet	Campylobacter jejuni plasmid pTet	Campylobacter hominis ATCC BAA-381, complete genome	Campylobacter coli plasmid pCC31	Campylobacter fetus subsp. fetus 82-40, complete genome	Clostridium difficile 630, complete genome	Campylobacter hominis ATCC BAA-381 plasmid pCH4	Campylobacter coli RM2228 plasmid pCC2228-3	Campylobacter jejuni RM1221, complete genome	Campylobacter jejuni plasmid pCJ1170	Helicobacter acinonychis str. Sheeba, complete genome	Campylobacter lari plasmid pCL300	Methanosarcina barkeri str. fusaro chromosome 1, complete genome	Helicobacter pylori plasmid pHel5	Homo sapiens chromosome 20, reference assembly, complete sequence
	NC_006270	**NC_004949**	**NC_008790**	**NC_007141**	**NC_006135**	**NC_009714**	**NC_006134**	**NC_008599**	**NC_009089**	**NC_009713**	**NC_008051**	**NC_003912**	**NC_008052**	**NC_008229**	**NC_006975**	**NC_007355**	**NC_004949**	**NC_000020**
Contig914 2794	0.007;52 94.12																	
Contig875 2110		0.006;52 96.67																
Contig1023 3709			3e-06;64 88.33	3e-06;64 88.33	3e-06;64 88.33	1e-05;62 90.20	1e-05;62 87.30											
Contig1031 3394								0.0;751 89.74										
Contig1042 4248			0.0;5923 97.38	0.0;3856 97.12	0.0;5919 97.35		0.0;3903 97.56		0.003;54 96.77									
Contig1030 2579								1e-56;230 100.00										
Contig872 2073			0.0;2083 98.64	0.0;2083 98.64	0.0;2083 98.64		0.0;1437 97.49											
Contig1181 8733								0.0;749 99.23										
Contig1165 8984			4e-10;78 82.22	4e-10;78 82.22	4e-10;78 82.22	7e-15;94 85.59	2e-06;66 81.02											
Contig1185 5740										6e-42;183 92.42	3e-26;129 81.02		2e-23;12: 82.11	7e-11;80 85.87	7e-08;70 90.91	2e-05;62 89.09	0.015;52 96.67	0.015;52 100.00
Contig1139 3987								0.0;735 96.88										
Contig1186 9928								7e-09;74 100.00										
Contig846 3329			0.0;809 96.89	0.0;809 96.89	0.0;809 96.89		1e-162;581 90.60					0.0;2220 94.68						

***Complete Cfv specific Contig1018, Contig1021, Contig1024, Contig1042, Contig1120, Contig419, Contig733, Contig851, Contig958, Contig991**

4.3.3 *Cfv* open reading frame analysis

The *C. fetus* subsp. *venerealis* 1474 ORFs protein database search found 67 unique to *Cfv* (no protein alignments), 1174 conserved top match alignment to *Cff*, 116 conserved top match alignment to any other species, and 117 low significance alignments. ORF alignments to the non-redundant protein database found 12% *Cfv* insignificant and unique (Appendix 4.2), 51% with significant alignments and 37% with highly significant alignments. Comparison of the 9 *Campylobacter* genome protein datasets found approximately 50% of proteins were in common for all *C. jejuni* (including subsp. *jejuni* and *doyley*) except *C. jejuni* subsp. *jejuni* 81-116 which had 20-25% similar. This level of similarity was also found between the *Cff* subspecies while between all *Campylobacter* species this similarity decreased to between 0.5-5.5%. The BlastMatrix [294] result can be found in Appendix 4.5

4.3.4 *Cfv* Open reading frame analysis of the *Cfv* specific suite of genomic regions

Eighteen *Cfv* specific contig ORFs were analysed against all available protein datasets. These *Cfv* specific regions contained 90 ORFs, 15 with alignments to hypothetical proteins, 32 with non-significant protein alignments and 43 ORFs with significant alignments. As a separate category these latter 43 ORFs were found to have significant alignments to plasmid/phage like proteins within *Campylobacter* species (34 ORFs) and to other bacteria (9 ORFs). In the 34 *Campylobacter* ORFs, best matches were found in two *Cfv* ORFs, namely a putative type IV secretion

system protein identified in Is*Cfe*1 [288] and a putative TrbL/VirB6 plasmid conjugal transfer protein. The remaining 32 ORFs had significant hits to *Campylobacter* species other than *Cfv* such as *C. curvus* (1), *C. concisus* (2), *C. coli* (4), *C. fetus* (5), *C. jejuni* (13) and *C. hominis* (17).

Functional assignments for the *Cfv* specific ORFs were as follows: cellular processes and signalling, chromosome partitioning, cell motility and intracellular trafficking, secretion and vesicular transport (16); information storage and processing, replication, recombination and repair, transcription, translation (12); metabolism and transport amino acid, carbohydrate and inorganic ion, energy production and conversion (5); and poorly characterized, general function prediction only (7) (Appendix 4.2).

4.3.5 *Cfv* IS*Cfe*1 insertion elements

Specific sites previously identified for the IS*Cfe*1 insertion element [288] were searched in *Cfv* alignments to *Cff* (Figure 4.2) : (a) the sodium／hydrogen exchanger protein gene *nahE* (YP_891382) was found in the *Cfv* pseudomolecule positioned 159601-160764 bp (Contig1097), a region conserved with *Cff*; (b) the putative methyltransferase protein gene metT (YP_892765) was found in the *Cfv* pseudomolecule positioned 1,605,092-1,603,530 bp (Contig1155) a region also conserved in *Cff*; and (c) the putative VirB6 protein gene was found in a number of *Cfv* contigs, these include contigs with ORFs not in common with *Cff* Contig1023 and *Cfv* specific Contig1165, Contig733, Contig875 and Contig958.

Cfv contigs were searched for the IS*Cfe*1 insertion containing sequences (AM260752, AM286430, AM286431 and AM286432). All the IS*Cfe*1 sequences aligned to Contig993 (39-1464 bp) with greater than 90 percent identity. Contig993 *Cff* position is indicated in Figure 4.2. The IS*Cfe*1 genes *tnpA* and *tnpB* matched Contig993 orf1 partial transposase A (*Cfv*) (14-157) and Contig993 orf2 transposase B (*Cfv*) (144-1436). Upstream ORF regions in Contig993 are: Contig993 orf3 anaerobic C4-dicarboxylate transporter (*Cff*) 1509-1697bp; Contig993 orf4, anaerobic C4-dicarboxylate transporter (*Cff*) 1705-2493bp; and Contig993 orf6 which had no protein alignments 2795-2968bp.

Contig875 only aligned with AM286432 (21-1235 bp) putative virulence genes with less than 90% sequence identity. Contig875 orf3 (499-1068 bp) partially to the partial putative virulence gene VirB5 and Contig875 orf5 (1302-2069 bp) to the truncated putative TrbL/VirB6 plasmid conjugal transfer (*Cfv*) gene. Downstream in Contig875 were Contig875 orf1 transposase OrfA (*Helicobacter pylori*) 30-170 bp and Contig875 orf2 (274-489 bp) with no protein alignments.

4.3.6 Genomic plasmid analysis

Plasmid-containing *Campylobacter* includes *C. coli*, *C. lari, C. concisus* 13826 (2 plasmids), *C. hominis* ATCC BAA-381 (1 plasmid), *C. jejuni* subsp. *jejuni* 81-176 (2 plasmids) and *C. fetus* subsp *venerealis* strain 4111/108. Complete plasmids have been sequenced for *C. coli* (6), *C. lari* (2), other *C. jejuni strains* (6) and *C. fetus* subsp *venerealis* (1). A direct search of these extrachromosomal *Campylobacter* plasmid sequences against *Cfv* specific sequence determined plasmid borne

genes in common between the species. Plasmid sequences from *C. coli, C. hominus* and *C. jejuni* represent over a third of the *Cfv* specific ORFs (37/90). These include type IV secretion system (Vir and Cmg), ParA, Ssb, RepE, mobilization and plasmid (Cpp and pTet) proteins (Table 4.3). Transposase genes were absent in the other *Campylobacter* spp. plasmids and found in *Cfv* Contigs1185 (2), Contig872 (1) and Contig875 (1). The *C. fetus* subsp. *venerealis* plasmid *pCFV*108 (EF050075*)* contains four genes, putative *mobC,* putative *mobA, repE* and an uncharacterised orf3 [297]. Plasmid pCFV108 was not found in the *Cfv* contigs. A protein search however found significant alignments for Contig1185.orf00004 to MobA (ABK41363 489 aa) and Contig1185.orf00007 to RepE (ABK41364 351 aa) (Figure 4.3)

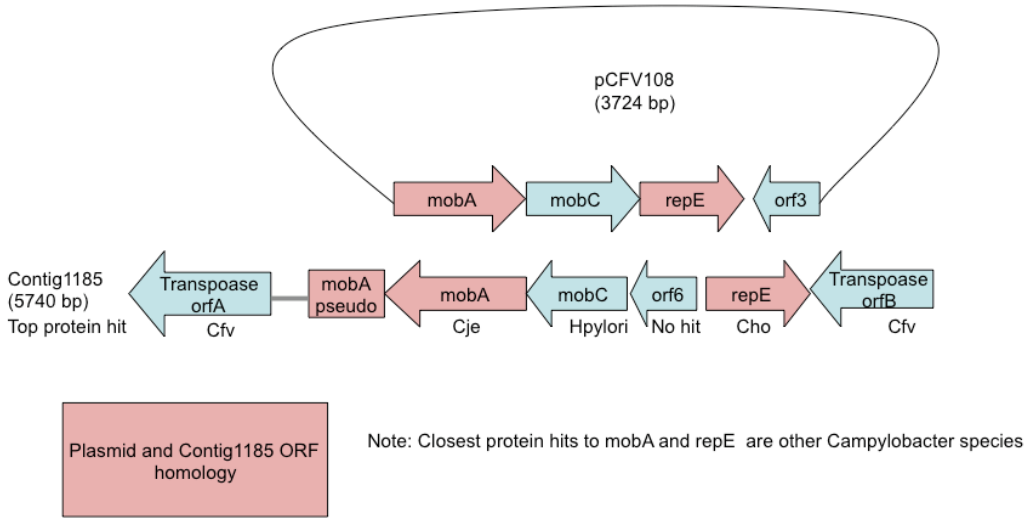


Figure 4.3 Plasmid pCFV108 insertion elements and Contig1185 ORFs with CFV protein best alignments (pink).

Table 4.3 *Campylobacter* plasmid gene comparisons. Comparison of all available *Campylobacter* plasmid gene content to the *Cfv* ORFs in the 80Kb *Cfv* specific suite of contigs. Plasmid-like genes are found in common between the different *Campylobacter* plasmid genes and *Cfv* specific ORFs.

Count of product	**Plasmid species**																				
Product	NC_004997 jejuni	NC_005012 jejuni subsp.	NC_006134 coli	NC_006135 jejuni subsp.	NC_006975 lari	NC_007141 jejuni subsp.	NC_007142 coli	NC_007143 coli	NC_007962 lari	NC_008049coli	NC_008050 coli	NC_008051 coli	NC_008052 jejuni	NC_008438 jejuni	NC_008770 jejuni subsp.	NC_008790 jejuni subsp.	NC_009713 hominis ATCC BAA-381	NC_009795 concisus	NC_009796 concisus	C. fetus venerealis	Grand Total
alpha-amylase																				1	1
bacteriocin-type signal sequence domain protein																			2		2
conjugative transfer regulon protein															1						1
DNA integration/recombination/inversion protein																				1	1
DNA topoisomerase I															1						1
DNA topoisomerase III																1					1
filamentous haemagglutinin family domain protein																		1			1
HmcD domain protein																			1		1
hypothetical protein	1	43			2	44		2	1					2	41	10	2	26	15	11	200
mobilization protein	1				1		1		1		1	1	1				2		2	1	12
NoProteinHit																				32	32
ParA															1			1		1	3
peptidase family protein																			1		1
RepA	1		1	1			1		1		1	1	1	1		1					10
RepB	1				1		1		1					1							5
replication initiation protein								1		1	1	1	1								5

site-specific recombinase, phage integrase family																			1			1
Ssb		1	1	1											1	1					1	6
tetO			1													1						2
TopA		1																				1
transcriptional activator																					1	1
TrbM protein																					1	1
vapD			1	1												1						3
phage portal protein																					1	1
RepE															1		1				1	3
pNL871104_p1																				1		1
periplasmic protein																			1			1
phytanoyl-CoA dioxygenase																			1			1
transposase																					2	2
DNA primase,																					1	1
stage V sporulation protein																					2	2
transposase OrfA																					2	2
type IV secretion system		7	10	10											5	9					14	55
phage			36	36											2	28			1	1	15	119
HipA																			1		2	3
Grand Total	4	52	50	49	4	44	3	3	4	1	3	3	3	4	53	52	5	33	23	90	483	

4.3.7 COG Analysis -Virulence Genes

The STRING database analyses identified 1141 *Cfv* ORFs that aligned significantly to STRING assigned COG functions. Comparative analysis between *Cfv* to the Cluster Orthologous groups found: 273 ORF in cellular processing and signalling a COG role known to contain virulence determinants; 164 information storage and processing; 406 metabolism; 153 poorly characterised; 87 hypothetical proteins; and the remaining without assignments to COG roles.

COG role distributions for virulence ORFs can be found in Appendix 4.4.

In putative virulence roles, 49 Cfv ORFs are involved in cell motility, 83 in cell wall/membrane/envelope biogenesis, 21 defence mechanisms, 25 intracellular trafficking, secretion and vesicular transport and 29 signal transduction mechanisms.

To identify virulence genes unique to *Cfv* or other *Campylobacter* species and distinguish the two subspecies, the *Cff* and *Cfv* virulence genes and *Cfv* contigs were aligned to the *Cff* genome. The non-redundant protein search also identified virulence genes such as Cytolethal distending toxin proteins (Cdt) [150], currently characterised as a hypothetical protein within the non-supervised orthologous groupings of STRING, although previously characterized and reported [298, 299].

Based on COG analyses (Appendix 4.4) the following sequences were selected for PCR validation. Those selected for PCR included virulence genes (including the Type IV secretion genes specific to *Cfv*) (8), flagella (6), cytolethal distending toxin (3), response regulator-sensor (6), membrane (4), fibronectin (1), haemolysin (1), Fe ABC transporter (1) and mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase (1) genes (Table 4.4).

Table 4.4 *C. fetus* validated primer set sequences and amplification results using *C. fetus* reference panel.

***Cfv* Contig gene ID**	**Gene ID (assay specificity[1])**	**Primer label**	**Primer sequence 5' to 3'**	**Anneal °C**	**Amplicon (bp)**	***Cfv* reference strains**					***Cff***	
						19438 **biovar ven**	98-109383 **biovar ven**	AZUL-94 **biovar ven**	**Pfizer biovar ven**	**Pfizer biovar interm**	15296	98- 118432
733 orf 1	*virB4* (*Cfv* AZUL-94 strain specific)	C733G3F C733G3R	TGATAAATAAAGAACCTGTTT TTTTTGCATAATCATTGTT	45	516	-	-	+	-	-	-	-
875 orf 5	*virB6* (*Cfv* biovar *venerealis* specific)	n875g2F n875g2R	GTGAATACACTTCAATCCG CAGTTTCCAGCATTCATAG	49	185	+	+	+	+	-	-	-
878 orf 2	*Flagellar flgH*	C878g1F C878g1R	AGAAAACGGCATAGGCGTAA AGTGCCGCTTCCGCTATAAT	55	301	+	+	+	+	+	+	+
927 orf 1	*Cytolethal distending toxin cdt*A	NC927g1F N927g1R	TCGCACGTAGTATGAACAAC CGTATAATCCTGTTCCGGTA	45	268	+	+	+	+	+	+	+
927 orf 2	*Cytolethal distending toxin cdt*C	NC927g2F NC927g2R	GCAACAGCTTCTATCTGAACAG ATCCTTTTGGAACCGTGC	45	175	+	+	+	+	+	+	+
927 orf 3	*Cytolethal distending toxin cdt*B	NC927g3F NC927g3R	GAGCGTTTGGAGCGATAA GCGGCCATAGTAGAAAATGT	50	154	+	+	+	+	+	+	+
988 orf 1	*Outer membrane protein (ompA)*	988G1F 988G1R	TTGAAGGAAACTGTGACGAGT CATACAGGTTTGCTCTCGC	50	150	+	+	+	+	+	+	+
992 orf 7	*Fibronectin Fn3*	992G5F 992G5R	GTTTTGGACTTTCTAATCCG AACGACATCTGTTCAGTATGAT	50	100	+	+	+	+	+	+	+
995 orf 5	Response regulator	nC995g4F nC995g4R	TCTTGATAAACTGCATAGCGCC CGCTGCTAAATGGACTTGAGAT	47	138	+	+	+	+	+	+	+
995 orf 6	Sensor	C995g5F C995g5R	TCCTTAGCTCAAATATAGTAGGATT AGGAAGTGGAATAGGTTTGTAT	47	107	+	+	+	+	+	+	+

1006 orf 4	Membrane protein	C1006G4F C1006G4R	AAGTATGGCAAACGGCG GTACGCTAATCTGTCGACTCTC	50	220	+	+	+	+	+	+	+
1013 orf 3	*Flagellar flhF*	nC1013g1F nC1013g1R	GCTTTCTAAACTTTCGCTTC ATATGCCCGCTTCTATGA	47	101	+	+	+	+	+	+	+
1023 orf 2	*virB10* (*Cfv* AZUL-94 strain specific	nC1023g1F nC1023g1R	AGTGGTGGATTTAAAAGCGGAC GTGGTAATCAACCCATCCTTCT	54	159	-	-	+	-	-	-	-
1023.orf 3 (NESTED)	*virB11* (*Cfv* AZUL-94 and *Cfv* biovar *intermedius* specific)	C1023G3F C1023G3R	ATATCAATGGAGTCTGGCAC AATGTTGTCTTACCACTGCC	55	349	-	-	+	-	+	-	-
1023 orf 3	*virB11* (*Cfv* AZUL-94 and *Cfv* biovar *intermedius* specific)	Nc1023g4F NC1023G4R	ACGCTGGTAGCGTAAAGCA CAACAACCTGTCTTTTGGCTC	55	161	-	-	+	-	+	-	-
1034 orf 10	Sensor	nC1034g7F nC1034g7R	GCCCATACCGAATTTTTCT ACGCCGATATATTTTTACTGG	45	101	+	+	+	+	+	+	+
1034 orf 12	Response regulator (*OmpR*)	C1034g9F C1034g9R	TTTGGGTATTTTGGATCATC AGCGGATCTAAAAGATAGAAGT	45	81	+	+	+	+	+	+	+
1037 orf 1	haemolysin secretion/activation protein, ShlB/FhaC/HecB family	C1037g1F C1037g1R	GCGATGAATATACCGTTAGAGG GAATAGTCTCGCTCGGCAT	50	291	+	+	+	+	+	+	+
1040 orf 1	Sensor histidine kinase	C1040g1F C1040g1R	ATAAGCCTACTAATCCCATCA AATGCTGCTTTTACCCAT	48	132	+	+	+	+	+	+	+
1047 orf 2	sensor histidine kinase	nC1047g1F nC1047g1R	AGGCGAGATCTTGGATCT GCGTGAGGACTTTGTGTTC	45	207	+	+	+	+	+	+	+
1083 orf 2	sensor histidine kinase	nC1083g1F nC1083g1R	TACAAATCACAAGACTACG CGAACTTACTATGAGT	50	102	+	+	+	+	+	+	+
1095 orf 4	iron uptake ABC transport	NC1095g2F NC1095g2R	AAAGTCTCTCATCAGTCG TACGCTCTTGATAGTGGT	55	250	+	+	+	+	+	+	+
1120 orf 4	*virB4* (*Cfv* biovar *venerealis* specific)	C1120G2F C1120G2R	TTCTCCTGCAACTGACGC GCTTTAACACGTCCGCC	50	521	+	+	+	+	-	-	-

1143 orf 3	*omp*	C1143G5F C1143G5R	GGCTTTAGAGGTACGGCTCC TAACGGACGTATCATACGCG	57	150	+	+	+	+	+	+	+
1154.orf 3	mannose-1-phosphate guanylyltransferase/ mannose-6-phosphate isomerase	C1154g3F C1154g3R	AAAAGCTGCAGTAGAGTTGG TCATCGAAACTTCCCATATC	55	429	+	+	+	+	+	+	+
1155 orf 4	*flaB* (*Cfv* AZUL-94, *Cfv* biovar *intermedius*, and *Cff* specific)	C1155g3F C1155g3R	ACTACCGCTTTGAGCAAGGA GGCGGTGCTAACGTATCATT	45	492	-	-	+	-	+	+	+
1165 orf 2	*virB9* (*Cfv* AZUL-94, *Cfv* biovar *intermedius*, and *Cff* specific)	nC1165g2F nC1165g2R	TGACAAAGATGAGCGGATAG TACCTGTTCGCCGTTTTC	50	151	-	-	+	-	+	+	+
1165 orf4	*virB11* (*Cfv* biovar *venerealis* specific)	nC1165g4F nC1165g4R	AGGACACAAATGGTAACTGG GATTGTATAGCGGACTTTGC	57	233	+	+	+	+	-	-	-
1165 orf 8	*virD4*(*Cfv* biovar *venerealis* specific)	nC1165g6F nC1165g6R	ATGTTCTAGCAGAGCTTGG TGACATTACGCCACTCTT	50	101	+	+	+	+	-	-	-
1172 orf 10	*Flagellar fliH*	nC1172g7F nC1172g7R	GCTTAAAACTATAACTCCGCCG TGCTAAAAGCTTGATCAGCG	47	160	+	+	+	+	+	+	+
C. fetus subsp. *fetus*	*Flagellar flhA*	nCFFFlhAF nCFFFlhAR	TTAAGCGAAGGCCATAATGG GTTTTCCAGGCATAGCATCA	50	202	+	+	+	+	+	+	+
C. fetus subsp. *fetus*	*Flagella flhB*	C999F C999R	CTGCGGTAGGGATATTTTGC TCCACTCAATGCTTCAGACG	45	543	+	+	+	+	+	+	+

[1]Assay specificity descriptions when not all reference strains are positive. Note 1023.orf3 has two PCR tests.

4.3.8 PCR diagnostics based on sequence identified in *Cfv*

To validate the subspecies specificity of virulence genes and *Cfv* specific sequences identified above, 31*Cfv* ORF sequences were selected in *Cfv* and primer sets tested using *Cff* and *Cfv* isolates (Table 4.4). Reference and type strains screened are described in Table 4.5 and *Cfv* reference strains included 4 *Cfv* biovar *venerealis* isolates (DPI, ATCC [300], UNSAM and Pfizer) and a *Cfv* biovar *intermedius* (Pfizer) isolate. *Cff* strains used were DPI and ATCC isolates as described in Table 4.5. All primers were based on the *Cfv* biovar *venerealis* AZUL-94 strain contig sequences except for *flhA* and *flhB* which were based on *Cff* sequence for these 2 flagella genes not identified in *Cfv* contigs. Conserved amplification of virulence genes in both *C. fetus* subspecies included flagella, outer membrane proteins, 2 component systems (response regulators and sensors), haemolysin, iron uptake and a fibronectin type III domain protein (Table 4.4). For assays based on ORFs selected as absent in *Cff*, contigs 1120 orf4, 1165 orfs 4, 8 and 875 orf5 assays amplified the *Cfv* biovar *venerealis* strains but not *Cfv* biovar *intermedius* or the *Cff* reference strains. These contigs were identified as: VirB4, VirB11, VirD4 and VirB6 type IV secretion system proteins respectively. Three assays (1023 orf2/VirB10, 1023 orf3/VirB11 and 733 orf1/VirB4) were specific for *Cfv* biovar *venerealis* AZUL-94 strain and did not amplify other biovar *venerealis* strains. One of these assays Contig 1023 orf3 (VirB11) also amplified *Cfv* biovar *intermedius*. *Cfv* biovar *intermedius* was negative in all other '*Cfv*' specific assays, which in the current study appear to be specific for *Cfv* biovar *venerealis*. Curiously, an assay based on 1165 orf2 (*Cfv VirB9*) was positive for *Cfv* biovar

venerealis AZUL-94, *Cfv* biovar *intermedius* and both *Cff* strains tested but did not amplify the other 3 *Cfv* biovar *venerealis* strains including the ATCC 19438 strain. All assays were specific for *C. fetus* subspecies, testing negative in related strains and reproductive disease pathogens listed in Table 4.5 including: *C. coli*, *C. jejuni, C. sputorum subsp. bubulus, C. hyointestinalis, Pseudomonas aeruginosa, Proteus vulgaris, Neospora caninum and Tritrichomonas foetus* (results not presented). However no single assay amplified all *Cfv* strains inclusive of both biovars *venerealis* and *intermedius*. Figure 4.4 demonstrates the specificity of selected primer sets Contig1023 orf2 and orf3, Contig1154 orf3 and Contig1165 orf4. Contig1023 orf3 and Contig1165 orf4 primers amplified sequences specific for *Cfv*, while Contig1154 orf3 primers amplified sequences in both *Cfv* and *Cff* strains.

Table 4.5 Reference strains tested in *C. fetus* PCR assays

Species and subspecies	Strain	Source[1]
C. fetus subsp. *venerealis*	98-109383 (Biovar *venerealis*)	Field Isolate (DPI&F, QLD)
C. fetus subsp. *venerealis*	19438 (Biovar *venerealis*)	ATCC 19438
C. fetus subsp. *venerealis*	AZUL-94 (Biovar *venerealis*)	UNSAM, Argentina
C. fetus subsp. *venerealis*	Biovar *venerealis*	Pfizer Animal Health
C. fetus subsp. *venerealis*	Biovar *intermedius*	Pfizer Animal Health
C. fetus subsp. *fetus*	98- 118432	Field Isolate (DPI&F, QLD)
C. fetus subsp. *fetus*	15296	ATCC 15296
C. coli	11353	NTCC
C. jejuni subsp. *jejuni*	11168	NTCC
C. hyointestinalis	N3145	Field Isolate (DPI&F, QLD)
C. sputorum subsp. *bubulus*	Y4291-1	Field Isolate (DPI&F, QLD)
Pseudomonas aeruginosa	27853	ATCC
Proteus vulgaris	6380	ATCC
Neospora caninum	50843	ATCC
Tritrichomonas foetus	YVL-W	Field Isolate (DPI&F, QLD)

[1]Legend: ATCC – American Type Culture Collection; NTCC – National Type Culture Collection; UNSAM – Universidad Nacional de General San Martín; DPI&F – Department of Primary Industries and Fisheries

Figure 4.4 PCR assay specificity for *C. fetus* subspecies and *C. fetus* subsp *venerealis* Examples of PCR assay specificity for *C. fetus* subspecies and *C. fetus* subsp *venerealis* biovars (*venerealis* and *intermedius*). Lanes numbered 1-4, N and M represent: 1 *Cfv* biovar *venerealis* 19438 ATCC, 2 *Cfv* biovar *intermedius* (Pfizer strain), 3 *Cfv* Argentina AZUL-94 strain, 4 *Cff* 15296 ATCC, N= negative no template control and M=molecular weight marker 100bp ladder (Invitrogen). Results are shown for assays based on Contig1154 orf3 (429 bp), Contig 1165 orf4 (233 bp), Contig 1023 orf2 (159bp) and Contig1023 orf3 (349 bp).

4.4 Discussion

The available *Cfv* genomic sequence information was aligned to the complete *Cff* genome sequence 82-40 in order to identify targets for the diagnostics for detecting *Cfv*. Based on the genome size estimates of *Cfv* [276, 301] and the completed *Cff* genome size, it is estimated that approximately 72% of the *Cfv* genome has been sequenced (unpublished, Prof Daniel Sanchez, Universidad Nacional de San Martin, Argentina). The ordering of available genome segments generally aligned well with the *Cff* genome as shown in Figure 4.2 and made evident a suite of *Cfv*

specific contigs. This suite of contigs contained a large range of type IV secretion factors, and plasmid/phage like proteins. A number of potential virulence factors were clearly identified as shared between *Cfv* and *Cff*. These virulence factors include an outer surface array membrane protein, chemotaxis types, motility associated, regulatory and secretion systems. The existence of these classes of genes with roles in the infection process, but not showing sub species specificity, is consistent with a two-tier infection model. '*Surface/membrane components provide necessary (but not sufficient) structural components for attachment to host cells. Specific components that complete the features of the surface/membrane structures are required for infection'* [150]. Many genes involved in host colonization have been found conserved across the *Campylobacter* genus [150]. Variations that were species specific were evident for a lipo-oligosaccharide locus, a capsular (extracellular) polysaccharide locus, and a novel *Campylobacter* putative licABCD virulence locus (not found in available *Cfv*). These observations are consistent with the suggestions that interactions between a pathogen's surface-exposed proteins and host cells represent a pivotal step in pathogenesis and virulence [302]. In pathogens several of the key players are proteins involved in adhesion, invasion, secretion, signalling, annulling host responses, toxicity, motility and lipoproteins [303].

Motility and chemotaxis genes have been found conserved among related *Campylobacter* species with flagella implicated in adhesion, protein secretion, invasion and virulence in pathogenic *C. jejuni* [150, 304-307]. Biosynthesis of

flagella requires the involvement of more than 40 structural and regulatory proteins including a type III secretion system for flagellar assembly [305, 307-309]. The *Cff flhA* gene based on genome alignments was found to be absent in the available *Cfv* sequence contigs, and coincided with the ordered alignment gap/non-sequenced section relative to *Cff*. However, one chemotaxis regulatory protein campy.fasta.screen.Contig1091 orf6 appears to be absent in *Cff* (Appendix 4.2). A lower complement of homologues associated with motility in *Cff* (n=41) compared with the other *Campylobacter* spp. (n=55-66) [150] has been identified in this study. However, the analysis of the incomplete *Cfv* genome identified a higher number of homologues (n=46) than the total *Cff* sequence. PCR assays based on a subset of flagellar genes (*flgH*, *flhF*, *fliH*, *flhA* and *fhlB*), demonstrated conservation of these sequences at least among the members of our panel of *C. fetus* strains including both subspecies (although *flhA* could not be identified in the available *Cfv* contigs). An additional assay designed to amplify the *flaB* sequence of the *Cfv* AZUL-94 strain did not amplify other *Cfv* biovar *venerealis* strains but did amplify *Cfv intermedius* and the *Cff* isolates. It has not been confirmed if this is attributed to *flaB* sequence variation or an absence of the gene in different geographical *Cfv* biovar *venerealis* strains, this gene has been targeted however for genotyping studies in other *Campylobacte*r species [310]. This study does confirm that the complete *Cfv* genome may harbour more flagellar/motility homologues than *Cff*. Virulent *C. jejuni* harbours more flagellar genes than less virulent species *C. coli*, *C. lari* and *C. upsaliensis* [150].

Adherence of other *Campylobacter* species to gut epithelial cells is mediated by multiple adhesins including *cadF* (*Campylobacter* adhesion to fibronectin); [311], PEB1 protein (putative binding component of an ABC transporter), [312], *JlpA* (jejuni lipoprotein A), [313] and a 43-kDa major outer membrane protein [314], confirmed as conserved in *C. jejuni*, *C. lari*, *C. upsaliensis* and *C. coli* genomes [150]. *Cfv* homologues for PEB1 and fibronectin-binding (FN-binding) proteins were confirmed with the remaining 3 absent in the genome contigs currently available. However, only the PEB1 protein was identified in the complete *Cff* genome sequence 82-40. Fibronectin is known to enhance *C. fetus* attachment [315] however in the absence of an identified *C. fetus cadF* homologue, it appears that the adherence mechanisms in *C. fetus* may differ from other *Campylobacter* species. In the case of *C. fetus* subsp. *venerealis*, this is perhaps not surprising as *Cfv* colonise the genital tract and not the intestinal tract, thus perhaps novel adhesins will be identified with completion of a *Cfv* genome sequence.

Toxin sequences, two component regulatory systems, plasmids and type IV secretion systems have also been recognised as components in pathogenic *Campylobacter* spp. [150]. Three cytolethal distending toxin (*cdt*) subunits A, B and C are confirmed as conserved across the four *Campylobacter* species (*C. jejuni, C.lari, C. coli, C. upsaliensis*) and *C. fetus* [298, 299]. In addition, the presence of *cdt* genes is linked to *C. jejuni, C coli* and *C. fetus* pathogenesis, where *cdt* negative strains were found to be less efficient during adherence and invasion *in vitro* [298, 316]. A similar survey of *C. fetus* will assist to confirm if a *cdt* positive

result is associated with an increase in pathogenicity. Two-component regulatory, TCR, systems are commonly used by bacteria to respond to specific environmental signals such as temperature [146]. Five TCR systems (pairs of adjacent histidine kinase and response regulator genes) have been identified as conserved across *Campylobacter* species and confirmed in *C. fetus* subspecies.

The type IV secretory genes, which are possibly involved in conjugative plasmid transfer or the secretion of virulence factors [150, 288, 317], were absent in the *Cff* genome and unique to *Cfv*. A large proportion of *Cfv* subspecies specific ORFs (30%) were harboured in the *Cfv* contig specific regions. *C. upsaliensis* and *C. jejuni* are known to harbour plasmids and evidence does suggest that these plasmids can play a role in pathogenesis. One basic difference between the list of genes absent in *Cff* and present in *Cfv* is that many of them are in common to genes present on the plasmids of these related *Campylobacter*. The type IV secretion system is also found in *C. jejuni*, *C. lari* and *C. coli* plasmid sequence. The unique *Cfv* genome sequences also harboured many phage-like derived genes. The presence of type IV secretion system has also been described by Abril et al. in 2007 [288], of which the putative VirB6 protein gene was found to be truncated by the insertion element (IS*cfe*1). It is possible that contigs within this *Cfv* unique 80Kb suite of contigs represent a number of extrachromosomal DNA plasmids. A wider survey of *C. fetus* isolates and the presence of plasmids (type IV secretion systems) and phage genes will assist to confirm our observations.

The work presented in this chapter analysis has provided diagnostic markers to discriminate the *Campylobacter* subspecies *Cfv* and *Cff,* which can be investigated for more general applicability in field use. Most of the *Cfv* assays based on the incomplete AZUL-94 genome sequence, showed amplification preference for *Cfv* biovar *venerealis* strains. The *Cfv* biovar *intermedius* strains were negative in all but one assay, which was otherwise positive for *Cfv* AZUL-94 strain only. Curiously, one of the assays designed to *Cfv* AZUL-94 strain *virB9* (type IV Secretion gene) did not amplify other *Cfv* biovar *venerealis* isolates but did amplify biovar *intermedius* and the *Cff* strains tested here. However, as described above the *Cff* genome sequence (Strain 82-40) does not appear to have type IV secretion genes. A confounding factor in interpreting this data is that different *Cff* strains may also possess putative plasmid-borne genes and these may potentially be shared between subspecies and *Cfv* biovars. The *Cfv* AZUL-94 strain could also either consist of a mix of the 2 biovars or represent a novel strain of *Cfv*. However, assays based on putative plasmid-borne genes have previously demonstrated inconsistencies when applied for subspecies identification in some regions [289]. The *parA* (plasmid partitioning protein gene), [318] assay target is thought to be plasmid borne, however evidence for plasmids containing *parA* in *Cfv* has not been confirmed to date [289, 318]. Very little research has been undertaken to compare the *Cfv* biovars and the diagnostic targets reported here now need to be further tested in multiple field strains to assess the stability of these markers and therefore the genomic regions in *Cfv*. However, the results presented do suggest that the *Cfv* research community could benefit from the generation of full genome sequence

from both biovars as well as isolates from different geographical continents. Our results also demonstrated putative plasmid sequences are present in *Cfv*, absent in *Cff*, suggesting plasmid profiling and sequencing from *C. fetus* subspecies, biovars and strains will assist to confirm our findings.

4.5 Conclusion

The PCR assays have highlighted the complexity of virulence factor specificity within *C. fetus* subspecies and strains, which are probably due to plasmid borne gene elements. We found that most genes important for interactions between a pathogen's surface-exposed proteins and host cells that represent a pivotal step in pathogenesis and virulence were conserved in *C. fetus.* These genes, although important, did not differentiate the subspecies and therefore were not the virulence factors that determined specificity. Instead we found the suite of extrachromosomal type IV secretion system, T4SS, *vir* genes specific to the *Campylobacter fetus* subspecies *venerealis* biovar *venerealis* AZUL-94 were able to consistently discriminate the *C. fetus* subspecies *fetus* in our PCR assays. The results provide the basis for the immediate identification of additional diagnostic tools as *Cfv* genomes are sequenced to completion, to develop the definitive diagnostic tool for comprehensive *Campylobacter fetus* subspecies differentiation. An ARC linkage project (LP0883837) with industry partners Pfizer Australia and Gribbles Veterinary Pathology will now characterise and study Australian isolates of *Cfv*. This project will use an integrated genomics sequencing approach, comparative genomics plus molecular and phenotypic screening to improve the understanding of the biology of genital campylobacteriosis in beef cattle.

In summary the key findings from this chapter are: 1) the bioinformatics approach provided the direct comparison of a partially sequenced genome to another reference sequence to identify genome specific regions; 2) the direct comparison of *Cfv* to *Cff* genome sequence identified an 80Kb region for *Cfv* specific analysis, 3) functional analysis of the *Cfv* 80Kb sequence found *CFV* specific virulence genes for diagnostics testing.

5 Chapter Five - Predicting gene targets in complex genomes: *Rhipicephalus microplus* target gene predictions for parasite control

The contents of this chapter have been published [319]. Laboratory tools, chemistries carried out by QDEEDI, Ala Lew-Tabor and Jessica Morgan.

5.1 Introduction

Ticks are obligate ectoparasite bloodsuckers, divided into two families the Argasidae (soft tick) and Ixodidae (hard ticks), that can transmit a wide variety of pathogens both these direct and indirect effects respectively can be lethal [320]. Tick control methods rely heavily on chemicals (acaricides), most of which target the central nervous system [321]. Neural-specific gene sequences are the target of acaricide chemicals. Increasing acaricide resistance, the contamination of ecosystems and food by chemical residues has led to alternative methods for control [322, 323].

5.1.1 Functional roles of genes in ectoparasite required for feeding

Tick salivary glands produce a complex mixture of proteins with a variety of functions important for prolonged attachment and feeding to the host [324]. Salivary component functions include, a cement latex-like secretion for secure attachment, antigenic and non-antigenic proteins (lipo and glycoproteins), anticoagulants to inhibit host blood clotting, host anti-inflammatory agents, vasodilators, antihistamines and agents that suppress components of the host

immune system, and agents that reduce the sensitivity of the receptors in the host's skin [324-329]. The host releases histamine proteins at the site tick attachment; the tick then secretes a histamine binding protein, HBP, that binds to the free histamine, suppressing the effect of the host response [328, 329].

5.1.2 Genomic targets for tick control

In response to the disadvantages that come from the use of chemical products for tick control, research into immunological methods to control tick have been conducted worldwide and constitute an important biological alternative [330]. The inoculation of cattle with appropriate synthetic oligopeptides derived from different regions of a protein can induce a protective immune response against the tick [320]. Synthetic peptides as a vaccine are advantageous as the quality and stability is more assured for cattle inoculations [331]. The synthetic peptides of epitopes defined in the tick could elicit protective antibodies in the host [332].

A single vaccine against tick currently exists, the tick vaccine (TickGARDTM in Australia and GAVACTM in South America) is derived from the 'concealed' antigen, Bm86, a midgut membrane-bound protein of the cattle tick, *Rhipicephalus* (*Boophilus*) *microplus* (*Rmi*) [333]. Anti-ectoparasite vaccines developed that control tick targets such as an 'exposed' tick saliva antigen and cross-reacts with 'concealed' tick midgut antigens have a dual target approach [334]. This approach of targeting antigens that are normally concealed from the host, and therefore not subjected to immune selection pressure [333] has been adopted in the development of vaccines for other ectoparasites [330, 335] . Vaccine-induced

immunity is usually antibody dependent [336]. In *Rmi*, the Bm86-derived tick vaccine antibodies bind to antigenic epitopes on the midgut cells of the feeding tick, this causes damage and leakage of blood into the body cavity, killing the tick or reducing fecundity [337]. The Bm86 vaccine is however only effective against adult and not the immature stages of *Rmi* tick species [338].

Pharmacological action of tick saliva on sera [339, 340], and the immunological profiles of cattle infested with *Rmi* based on local reactions to tick bites have been previously reported [341, 342]. These new approaches to the development of vaccines [343] have been tested including immune responses against recombinant tick antigen, for the development of vaccine Bm95 [344, 345]. Purified extracts of cement proteins secreted by tick salivary glands, infiltrate the epidermal/dermal layers of the host skin [346], and have been shown to be immunogenic and immunoprotective [347]. Anchoring tick salivary anti-complement proteins to a membrane increases immunogenicity [348] and antigen-specific antibody response in vivo to tick blood-feeding [349].

The cattle tick, *Rmi*, is one of the most economically important ticks affecting the global cattle population [350]. *Rmi* and its associated pathogens can be transmitted to cattle and lead to severe agricultural losses in milk and beef production and restrict the movement of livestock [351]. The most affected regions of the world are tropical and sub-tropical countries including northern Australia, Mexico, South

America and South Africa, with threats to USA cattle populations at southern borders with Mexico [351].

The genome sizes of three species of ixodid ticks, *Ambylomma americanum* [352], *Boophilus* (*Rhipicephalus*) *microplus* and *Ixodes scapularis* (*Isc*) [353] have been estimated using DNA reassociating kinetics. The *Rmi* genome has an estimated size of 7.1 Gb, three times the size of the *Isc* genome (2.3Gb) [353, 354]. The *Rmi* genome is found to be composed of fold back (FB), highly repetitive (HR) and moderately repetitive (MR) elements, in the following proportion 0.82% FB, 31% HR, 38% MR, and 30% unique DNA, similar to *Isc* [353]. A short interspersed repetitive element (SINE) Ruka element, containing RNA polymerase III promoters, is major component of eukaryotic genomes that are particularly abundant in the heterochromatic compartment of vertebrates and plants as reviewed Kidwell and Sunter [355, 356]. SINE transposable elements have the ability to move to new locations based on reverse transcription prior to genomic integration. Most SINEs are derived from tRNA [357], although some, such as the Alu family which accounts for approximately 10% of the human genome, are thought to originate from 7SL RNA sequences [358] (see chapter 3). It has been shown in *R. appendiculatus* that secondary structure predictions indicate Ruka could adopt a tRNA structure similar to a serine tRNA [355].

The *Isc* Genome Project (IGP) [153, 359], is the first tick genome sequencing effort and currently a major resource for tick comparative genomic analyses. This project

has influenced the rapid rise in the number of sequences for tick DNA in NCBI [360]. The current *Isc* genome draft, represented by 369,492 supercontigs, (1.7 Gb) of linear genomic sequence, was used in this analysis to identify conservation with available *Rmi* genomic DNA.

To provide insights into the complexity of the tick genome and identify vaccine targets, the following *Rmi* sequence resources were available for analysis.

1) The BmiGI Version 2 gene index [361] containing 13,643 non-redundant tentative consensus transcript gene sequences.
2) *Rmi* Cot reassociating kinetics genomic sequence, that has been demonstrated as a useful tool to explore the gene space of large genome species [362].
3) A BAC end library, created with the view to probe the *Rmi* genome for BAC sequencing [363].
4) A suppressive subtractive hybridization (SSH) to identify transcripts associated with host attachment and/or feeding, which identified both a large increase in rRNA transcripts thought to be associated increase protein production during tick feeding, and the production of a number of enzymes including serine protease inhibitors (Serpins) [151].

5.2 Materials and Methods

5.2.1 BAC end sequences

Glycerol plates with BAC clones (1,125 96-well plates) were submitted to Beckman Coulter Genomics (Beverley, MA, USA) to obtain approximately 12,000 reads using bi-directional sequencing of the clones. The Beckman Coulter Genomics

protocol is described as follows: clones were picked from the 96 well plates, cultured and DNA was purified using SPRI®; following dye-terminator fluorescent sequencing the product was purified using CleanSEQ® with sequencing fragments detected via ABI3730xl capillary electrophoresis. The total 10,582 BAC end sequences (BES) provided as trace files from Agencourt were clipped of the vector (pECBAC1) with cross_match, Phrap package version 0.990329 [364]. Sequences greater than 500bp have been deposited GenBank GSS under HN108288-HN118367 [365].

5.2.2 BAC genomic DNA extraction, library construction, and BAC screening and sequencing

Ticks from the Deutsch strain of *R. microplus* were reared at the USDA-ARS Cattle Fever Tick Research Laboratory in Mission, TX [366]. Genomic DNA extraction, library construction, and BAC screening are as described by Guerrero et al., [362].

5.2.3 BAC sequencing

The BAC vector used was pECBAC1 and the cloning site BamHI. BAC libraries were sequenced using 3-4kb insert high copy shotgun library methods, aiming for 8-fold coverage of 1,008,000 bases (high copy) using Sanger Sequencing ABI technology (Beckman Coulter Genomics, MA, USA). Phred 20 read lengths were greater than 700 bp and had pass rates: > 90% and x 6 coverage.

5.2.4 BAC assembly

The Sanger BAC sequences were assembled with Phred/Phrap [162]. The following tools were also tested for assembly validation CAP3 [163] and Phusion

[164] and MIRA [165]. The BES were mapped to the assembly with BLAT [367] at 100% percent sequence identity. Dot plot matrices were generated using Dotter64 [368]. Beckman Coulter Genomics (MA, USA) closed the sequence gaps based on pair end read linkage. The sequencing of 5 of the BACs has been reported [362], and the sequences from another 10 BACs have been deposited in GenBank (Accessions HM748958-HM748967) [365]. This chapter focuses on BAC sequences, BM-005-G14 (HM748961) and BM-012-E08 (HM748964).

The correct orientation and ordering of the contigs was based on pair-end read assembly linkage results, BAC end sequence alignment positions and gene annotation, as comparative analysis of *Rmi papilin* to a number of species show ordered domain conservation. The final BAC sequence length with gap closure was 135Kb, close to the estimated restriction digest size. In BM-005-G14 a *papilin* gene was first predicted with Genscan HumanIso model from the 2 large coding sequences CDS8 and CDS6 (6,663 and 4,077 bp respectively), which covered all the *papilin* protein domains, except the initial 5' end thombospondin domain. In addition, CDS8 contained a helicase domain (not previously identified in *papilin*), an Adam-TS-spacer 1 and the second expected thombospondin domain. CDS6 contained all the remaining *papilin* domains, Kunitz (KU) domains x10, Whey Acid Protein (WAP), Immunoglobulin (IG) x3 set, and the final 3'end Cytoplasmic phospholipase A2, catalytic subunit (PLAC) domain. Direct cDNA sequencing confirmed the BAC data and 5' RACE assisted to confirm the presence of the missing thrombospondin domain and our model subsequently proposed the

presence of 2 genes, the *papilin* and the *helicase*. The *papilins'* coding region of 8Kbp spanned a genomic region of 86Kbp.

5.2.5 BES analyses

BlastN [33] nucleotide similarity searches were conducted on the Dana Faber Cancer Institute (DFCI) BmiGI Gene Indices [152, 361], the IscGI [369], subtraction library cDNA [151] and iscapularis.preliminary.TRANSCRIPTS_JCVI-IscaW1.0.5 and NCBI [370] datasets that included nucleotide (nr), EST, RefSeq, GSS, and WGS.

BlastX [33] protein similarity searches were conducted on NCBI [370] (nr, patent), and *Ixodes scapularis* peptide gene predictions 1.1 iscapularis.PEPTIDES-IscaW1.1 protein datasets. Domain and protein family identification was conducted with RPSBlast, on NCBI Conserved Domain Database (CDD) [148] database. Genome searches were conducted with BLAT [367] to compare to the ixodes_scapularis_supercontigs.

5.2.6 Gene prediction

Genscan [174] (model for human isoforms) was used to assemble the BAC contigs. Bioperl [371] scripts were used to parse alignments and identify conserved regions. The *in silico* workflow was designed based on open source applications and CCG Grid computing [372]. Appendix 5.1 contains AutoFACT [373] annotation for 15 *R. microplus* BAC sequences.

5.2.7 Sequence alignment and phylogeny

ClustalW [34] was used for multiple sequence alignments and multiple sequence alignments for the manuscript were displayed in Jalview [374].

Phylogeny analysis was conducted with the Phylip version 3.6 [375] protein distance algorithm and the Neighbor-Joining method [376] and a bootstrap test of 100 replicates.

The molecular clock test was performed by comparing the ML value for the given topology with and without the molecular clock constraints under the Jones-Taylor-Thornton (1992) model [377, 378]. Evolutionary analyses were conducted in MEGA4 [378]. The evolutionary history was inferred using the Neighbor-Joining method [376]. The bootstrap consensus tree inferred from 500 replicates was taken to represent the evolutionary history of the taxa analyzed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) is shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the molecular evolution and phylogenetics method [379] and are in the units of the number of amino acid differences per site. The analysis involved 7 amino acid sequences. All base positions containing gaps and missing data were eliminated. There were a total of 1124 base positions in the final dataset.

5.2.8 Repeat identification

Arthropod known repeats were identified with RepeatMasker version 3.2.6 [380]. Repeatscout [172] was used for the *de-novo* identification of repeat motifs. Perfect

tandem repeats were identified with a SSR finder Perl program written by S. Cartinhour, in 2000. Sine RUKA elements were identified based on BlastN [33] homology to GenBank: EU018139.1 (9,947-10,084 bp), percentage identity greater than 84% and coverage greater than 69%.

5.2.9 cDNA preparation

Total RNA was extracted from tick samples using the Trizol reagent purchased from Invitrogen Corporation, CA, USA. Tissue was ground to a fine powder using a mortar and pestle with liquid nitrogen and the powder transferred to a tube of Trizol with 1 mm glass beads. This mix was further homogenised for 45 seconds in a MiniBeadbeater-96 (Biospec Products, Bartlesville, OK, USA) then the RNA was extracted using chloroform and isopropanol. Double stranded cDNA was created from 25 µg of total RNA using a SuperScript™ Double-Stranded cDNA Synthesis Kit following Kit protocols (Invitrogen Corporation, CA, USA).

5.2.10 *Papilin* PCR amplification and sequencing

Primers based on BAC sequences were designed with EMBOSS [295] eprimer3 [296] and a minimum GC clamp of 2. Synthesis of primer sequences was by Sigma Aldrich (MO, USA) and sequences are presented in Appendix 5.2 tables 2 and 3. The full *papilin* cDNA was PCR amplified from cDNA extracted from frustrated larvae and cloned in three steps:

1) A 7723 bp product was amplified between primers Papilin57383F and PapilinR3 (designed from predicted coding sequence) using the Expand Long Template PCR System (Roche Applied Science, Mannheim, Germany) using Expand Long

Template Buffer 2. This reaction was thermocycled in a DNA Engine (PTC-200) Peltier Thermal Cycler (Biorad Laboratories, CA, USA). The purified product was transformed into chemically competent One Shot® TOP10 cells using a TOPO-XL ® PCR Cloning Kit (Invitrogen Corporation, CA, USA). For each transformation, DNA was prepared from six clones using a QIAprep Spin Miniprep Kit (Qiagen, CA, USA). Plasmid inserts were sequenced using Big Dye Vers 3.1 technology (Applied Biosystems, CA, USA) and were run on an Applied Biosystems 3130xl Genetic Analyser (Griffith University DNA Sequencing Facility, School of Biomolecular and Biomedical Science, Griffith University, Qld, Australia). Sequences were edited and aligned in Sequencher (Vers 4.8 Gene Codes Corporation, Ann Arbor, MI, USA). Additional sequencing primers were designed manually (Appendix 5.2).

2) The start codon for the *papilin* gene was determined following 5' amplification of the cDNA ends from larval cDNA using the SMARTer™ RACE cDNA Amplification Kit as described by the kit manufacturer (Clontech Laboratories Inc., CA, USA). The 5'-RACE PCR used an Advantage 2 PCR kit (Clontech Laboratories Inc., CA, USA) using the kit 5' RACE Universal Primer A Mix (UPM) primer and the gene specific reverse primer AdamSR1 designed within the Adam spacer region of the papilin gene (Appendix 5.2). The gel-purified product was cloned into chemically competent One Shot® TOP10 cells using a TOPO® TA Cloning Kit (Invitrogen Corporation, CA, USA). Clone inserts were sequenced as described above.

3) The *papilin* stop codon was determined from the predicted coding sequence and a primer was designed anchored at the stop position PapStopR1 (Appendix 5.2). A

611 bp product was PCR amplified between primers Pap12440F to PapStopR1. The product was cloned and sequenced as described above.

5.2.11 *Papilin* cloned products

The final *papilin* cDNA product was 8,761 bp. This was assembled from: 1) the 5' Race to the AdamS_R1 primer, product of length 867 bp; 2) the region between primers papilin57383F and PapilinR3, product length 7,723 bp and pap12440F to pap13230R direct sequence.

5.2.12 *Helicase* PCR amplification and sequencing

A second large PCR product 4886 bp in length was amplified from the larval cDNA between primers PapilinORFF2 and Papilin54900R. The 3' end of the product has a 229 bp overlap with the papilin gene (exon 2 and 3, 5' of the Adam spacer). The product was amplified using the Expand Long Template PCR System (Roche Applied Science, Mannheim, Germany) and was cloned and sequenced as described above for the large papilin clone. Internal primers were designed to sequence the complete clone that was found to contain a helicase gene (Appendix 5.2).

5.2.13 BM-012-E08 PCR

The PCR 50 μl reaction contained Advantage 2 SA PCR buffer, 10mM of dNTP mix, 10μM of each primer, 100ng of DNA template and Advantage 2 polymerase mix as recommended by the manufacturer (Clontech Laboratories Inc., CA, USA), Cycling Parameters (BIORAD DNA Engine Cycler). The initial denaturation was for 2 mins at 94^{o}C followed by 29 cycles of denaturation 1 min 94^{o}C, annealing 1 min

55°C, and extension 1 min 72°C, with a final extension of 7 mins 72°C. The products were visualised following agarose gel electrophoresis (1.2%) containing Gel Red (Jomar Bioscience Pty Ltd, SA, Australia). PCR products and purified plasmid DNA were sequenced using Big Dye Vers 3.1 technology (Applied Biosystems, CA, USA) and were run on an Applied Biosystems 3130xl Genetic Analyser at Griffith University DNA Sequencing Facility (GUDSF). Sequences were edited and aligned in Sequencher (Vers 4.7 Gene Codes Corporation).

Amplified products were cloned using the pCR2.1 – TOPO plasmid vector (Invitrogen Corporation, CA, USA). Transformed cells were plated on to LB agarose plates containing 50µg/ml kanamycin and grown overnight at 37°C. Colonies were picked and cultured in LB medium broth containing 50 µg/ml kanamycin. PCR reactions were performed on 1µl of the cultured broths and analysed by agarose gel electrophoresis to confirm insertion. Plasmid DNA was purified as described above. BAC BM-012-E08 primer sequences can be found in Appendix 5.2.

5.2.14 BM-012-E08 Long range PCR

Tick genomic DNA was prepared following tissue grinding as described above for cDNA preparation and subsequently purified using the QIAamp DNA mini-kit as described by the manufacture (QIAGEN, CA, USA). The Expand Long Template PCR system was used to amplify the DNA under conditions recommended by the manufacturer (Roche Applied Science, Mannheim, Germany) in a BioRad DNA Engine Peltier Thermal Cycler. Direct sequencing was undertaken as described

above. All sequence alignment graphs were generated using Bioperl Biographics modules [371].

5.2.15 qRT-PCR analysis

Primers were designed manually within targeted exon regions for the *helicase* and *papilin* transcripts described in this study (Appendix file 5.2). Methods for qRT-PCR analysis were described previously by Lew-Tabor et al. in 2010 [381], utilising tick extracts prepared from different tick organs and stages, normalised against 2 housekeeper genes (*Actin* and rRNA 18S) and against a pooled cDNA sample.

5.2.16 Cot selected genomic DNA

To enrich for single/low-copy and moderately repetitive DNAs, the Cot filtration of *Rmi* genomic DNA was performed as previously described [362]. The two "conditions" are called Cot69 and Cot696. Starting DNA concentrations for both were 200 micrograms of sheared genomic DNA. Time for renaturation was 1 hr, 48 min, 6 sec for sample Cot696 (Cot of 695.6) and 10 min 49 sec for sample Cot69 (Cot of 69.56). Renaturation was conducted at 70 degrees C, at 0.03 M NaPO4. Sequencing results form these two Cot-selected samples have been deposited in GenBank SRA, submission: SRA012677.4/SID00001.

Cot DNA 454 read sequence was mapped to the BAC sequence using the Newbler GS-FLX reference mapper, version 2.0.00.20 [166], and BLAT [367]. The total read number was counted for each BAC window size 100bps (read number per 100 bp window). The percentage (read number per 100 bp window) / (Total BAC read

count) was then plotted for each BAC. The gnuplots were calculated by (Sum bp depth / window size) / (Sum total bp depth / BAC length).

5.3 Results

5.3.1 Bioinformatics workflow

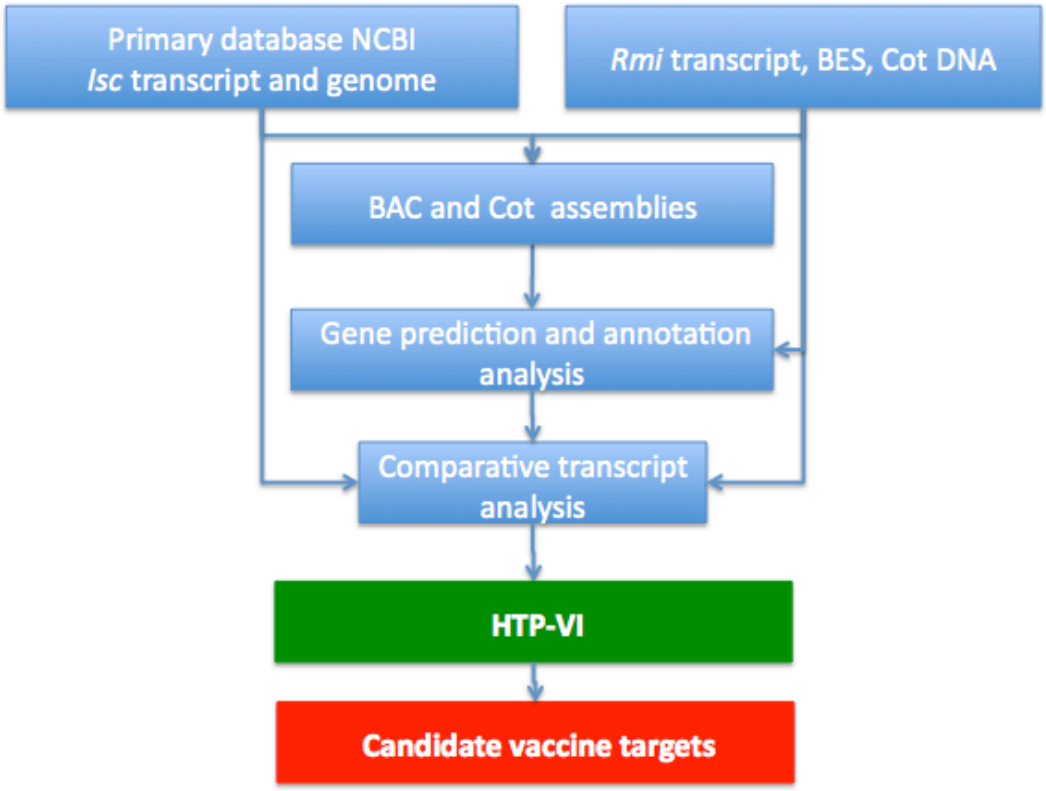


Figure 5.1 bioinformatics overview for the identification of genomic vaccine targets in *R. microplus.* BES = BAC end sequence, Cot DNA = sequenced DNA that has been filtered by reassociating method Concentration/time, HTP-VI = High Throughput Vaccine Identification workflow component.

Further detail for the High Throughput Vaccine Identification (HTP-VI) bioinformatics workflow, shown boxed in green in Figure 5.1, follows in section 5.3.7, and vaccine target (boxed in red) detail follows in section 5.3.8.

5.3.2 Genome sequence via BES and Cot DNA

Due to the large extent of repeats in *Rmi* genome (70%), a reassociation kinetics approach for partial genome sequencing was applied. Total *Rmi* genomic DNA was prepared and processed by two Cot filtration experiments to enrich for single/low-copy and moderately repetitive DNAs. Cot-filtered DNA was sequenced using 454 FLX pyrosequencing. Two Cot filtrations, Cot696 and Cot69, were enriched for single or low-copy (LR), and moderately repetitive (MR) DNAs respectively [362]. The aim was to isolate all genomic DNA that did not reassociate by Cot value of 696 M.s. and 69 M.s. The resulting ssDNA was used to make dsDNA by second strand synthesis and then digested to enrich for DNA fragments of sizes (250 to 600 bp) suitable for 454 FLX sequencing [362]. Starting DNA concentrations for both conditions were 200 micrograms of sheared genomic DNA. Time for renaturation was 1 hr, 48 min, 6 sec for sample Cot696 (Cot of 695.6) and 10 min 49 sec for sample Cot69 (Cot of 69.56). The two Cot short read data have been deposited in GenBank SRA, submission: SRA012677.4/SID00001. The two Cot filtrations were then used to estimate the frequency of a single full length *Rmi* RUKA element sequence in the genome to be at least 152,923 copies (discussed in further detail in section 5.3.8).

Over 10,582 BAC ends sequences (BES) (7,290,530 bp) were produced; this is the first BES project for *Rmi*. Those BES greater than 500bp have been deposited in GenBank GSS as accessions HN108288-HN118367.

The two Cot experiments and BES were then reassembled with a previous cot experiment [362], this Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank GenBank:ADMZ00000000 [382]. Sixty nine percent of the BES assembled with the Cot DNA.

The BES were comparatively analysed to the *Isc* representative genome comprising 369,492 *Isc* scaffolds (supercontigs), 58 *Rmi* BES aligned with greater than 80% BES coverage and 80% percent identity. The number of BES found in comparative databases searches at specified thresholds are, 2,418 (25%) at an expected alignment value (e-value) of 1e-05 to NCBI protein [370], 416 (3%) at an e-value 1e-20 to *Isc* proteins [97, 98], 2,559 (24%) BES at an e-value 1e-20 to *Rmi* gene indices [152], and 134 (1.2%) to *Isc* gene indices [369] at an e-value 1e-20.

5.3.3 BAC Analysis

In order to select BAC clones for sequencing, BAC end sequences (BES) were assessed against the complete protein resource at NCBI, the *Rmi* EST mixed library at the DFCI's Gene Indices, BmiGI [152, 361] accounting for approximately 14,000 tentative consensus transcripts [383]. This identified *Rmi* transcripts associated with host attachment and feeding of larval, adult female and adult male ticks from a suppressive subtractive hybridization (SSH) study [151]. Ten BAC clones were selected for sequencing and de-novo assembly [382], based on the above comparative analyses for hybridization to known transcripts of interest involved in tick feeding. These BACs have been deposited in GenBank (Accessions HM748958-HM748967) [382]. Figure 5.2 shows the graphs for 15

BAC sequences (10 BACs [365] and 5 BACs [362]) the Cot DNA frequency and GC content plotted over 1000bp window and 100bp step size. Figure 5.3 summarises the BAC sequence (x15) [362, 365] and gene predictions (CDS). The preliminary annotation can be found in Appendix 5.1.

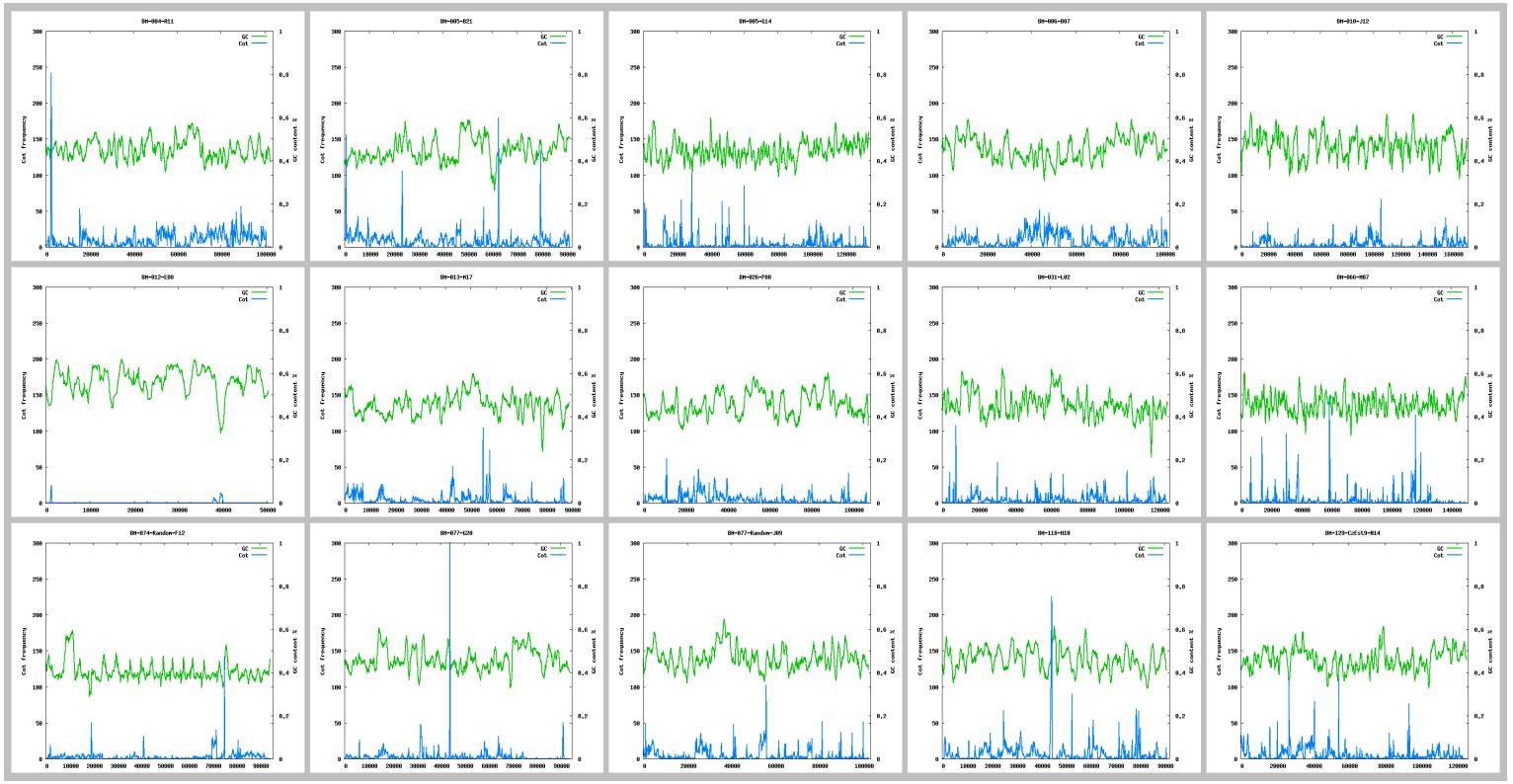
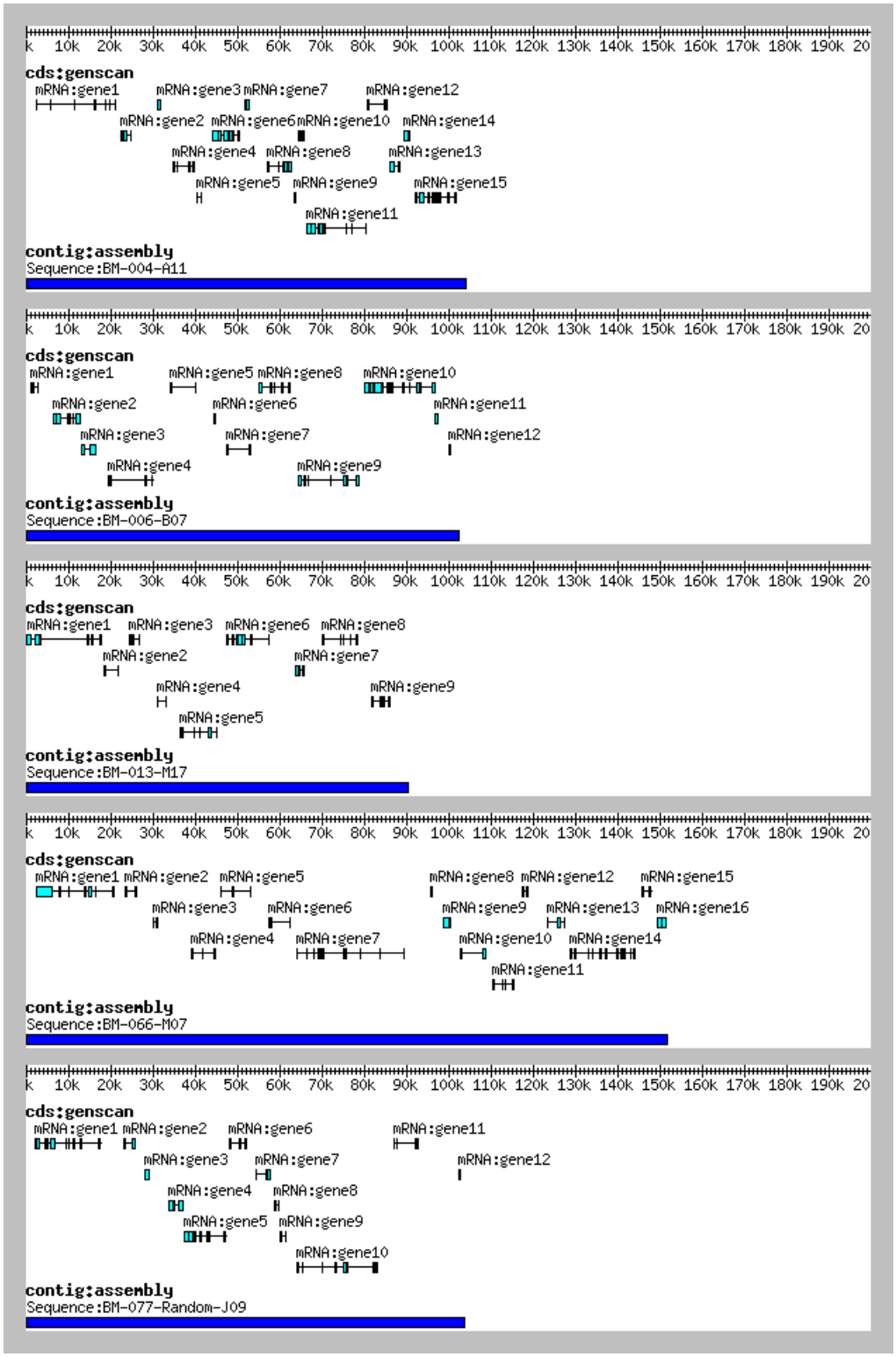
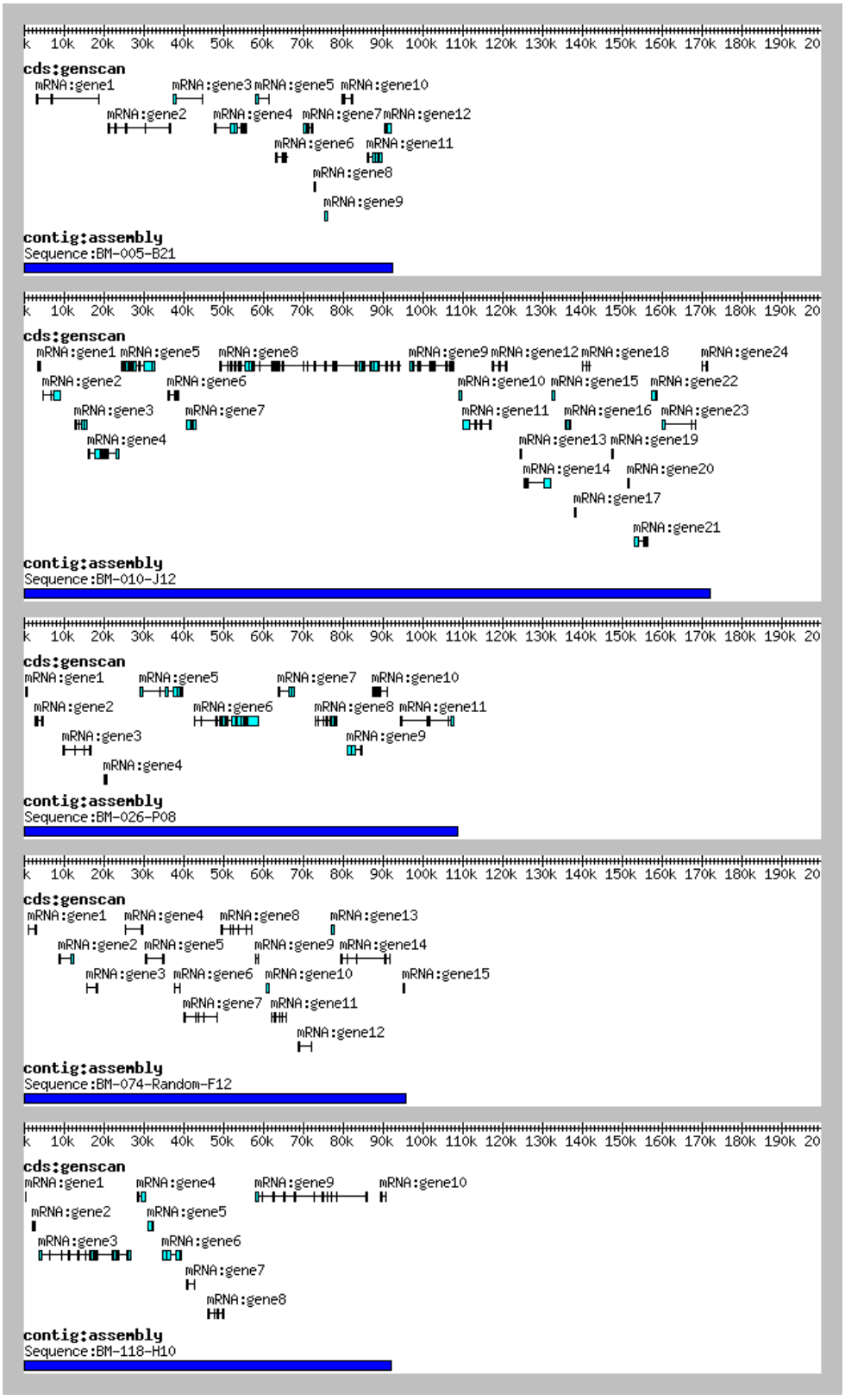


Figure 5.2 BAC sequence (x15) Cot DNA frequency (blue) and GC content (green) plotted over 1000bp window and 100bp step size


k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene3
mRNA:gene7
mRNA:gene12
mRNA:gene2
mRNA:gene6
mRNA:gene10
mRNA:gene14
mRNA:gene4
mRNA:gene8
mRNA:gene13
mRNA:gene5
mRNA:gene9
mRNA:gene15
mRNA:gene11
contig:assembly
Sequence:BM-004-A11
k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene5
mRNA:gene8
mRNA:gene10
mRNA:gene2
mRNA:gene6
mRNA:gene11
mRNA:gene3
mRNA:gene7
mRNA:gene12
mRNA:gene4
mRNA:gene9
contig:assembly
Sequence:BM-006-B07
k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene3
mRNA:gene6
mRNA:gene8
mRNA:gene2
mRNA:gene7
mRNA:gene4
mRNA:gene9
mRNA:gene5
contig:assembly
Sequence:BM-013-M17
k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene2
mRNA:gene5
mRNA:gene8
mRNA:gene12
mRNA:gene15
mRNA:gene3
mRNA:gene6
mRNA:gene9
mRNA:gene13
mRNA:gene16
mRNA:gene4
mRNA:gene7
mRNA:gene10
mRNA:gene14
mRNA:gene11
contig:assembly
Sequence:BM-066-M07
k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene2
mRNA:gene6
mRNA:gene11
mRNA:gene3
mRNA:gene7
mRNA:gene12
mRNA:gene4
mRNA:gene8
mRNA:gene5
mRNA:gene9
mRNA:gene10
contig:assembly
Sequence:BM-077-Random-J09


k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene3
mRNA:gene5
mRNA:gene10
mRNA:gene2
mRNA:gene4
mRNA:gene7
mRNA:gene12
mRNA:gene6
mRNA:gene11
mRNA:gene8
mRNA:gene9
contig:assembly
Sequence:BM-005-B21
k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene5
mRNA:gene8
mRNA:gene9
mRNA:gene12
mRNA:gene18
mRNA:gene24
mRNA:gene2
mRNA:gene6
mRNA:gene10
mRNA:gene15
mRNA:gene22
mRNA:gene3
mRNA:gene7
mRNA:gene11
mRNA:gene16
mRNA:gene23
mRNA:gene4
mRNA:gene13
mRNA:gene19
mRNA:gene14
mRNA:gene20
mRNA:gene17
mRNA:gene21
contig:assembly
Sequence:BM-010-J12
k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene5
mRNA:gene7
mRNA:gene10
mRNA:gene2
mRNA:gene6
mRNA:gene8
mRNA:gene11
mRNA:gene3
mRNA:gene9
mRNA:gene4
contig:assembly
Sequence:BM-026-P08
k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene4
mRNA:gene8
mRNA:gene13
mRNA:gene2
mRNA:gene5
mRNA:gene9
mRNA:gene14
mRNA:gene3
mRNA:gene6
mRNA:gene10
mRNA:gene15
mRNA:gene7
mRNA:gene11
mRNA:gene12
contig:assembly
Sequence:BM-074-Random-F12
k 10k 20k 30k 40k 50k 60k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 20
cds:genscan
mRNA:gene1
mRNA:gene4
mRNA:gene9
mRNA:gene10
mRNA:gene2
mRNA:gene5
mRNA:gene3
mRNA:gene6
mRNA:gene7
mRNA:gene8
contig:assembly
Sequence:BM-118-H10

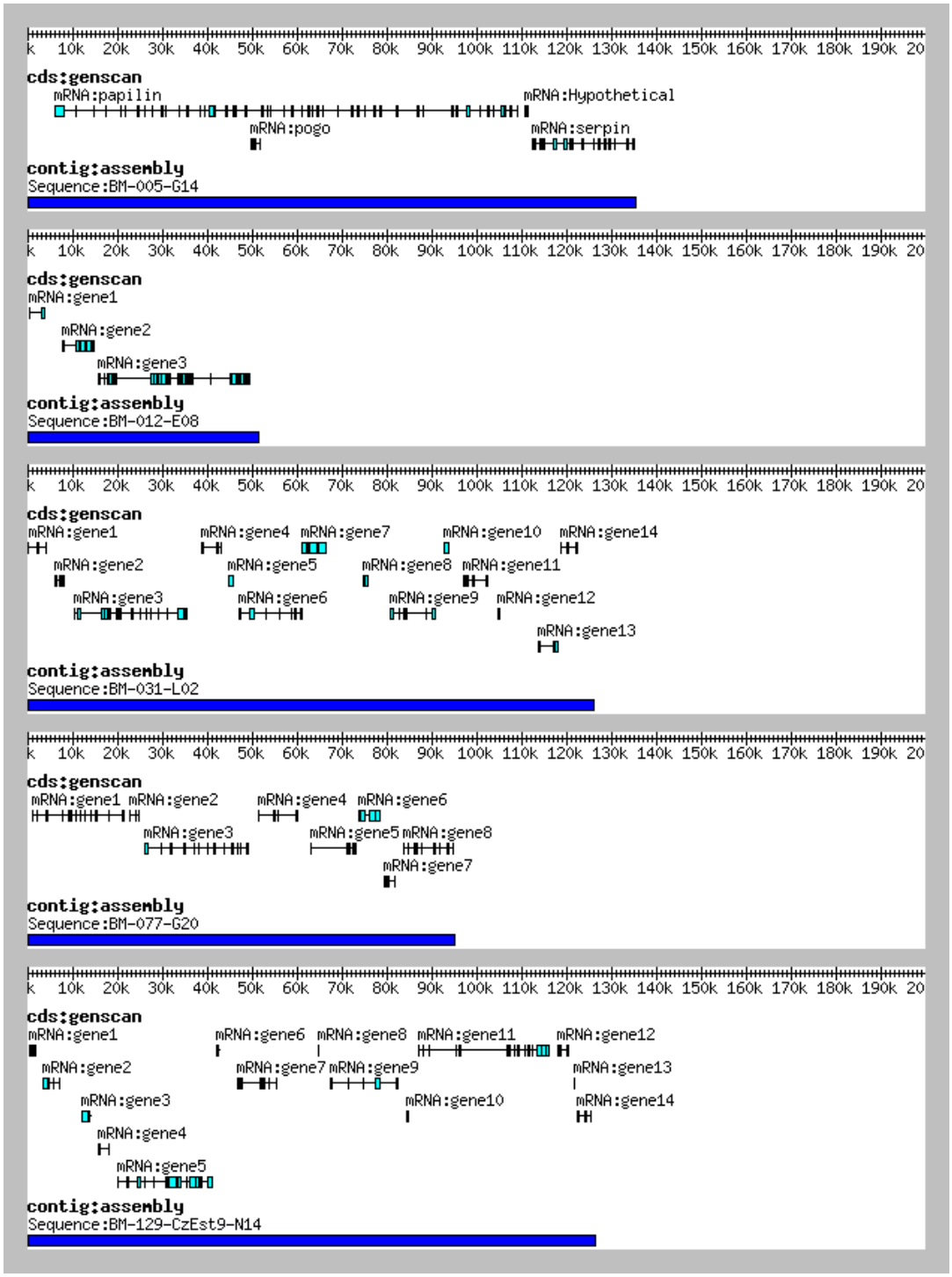


Figure 5.3 Gene predictions and preliminary annotation for 15 *Rmi* BAC sequences. Shown for each BAC are, mRNA CDS as light blue glyphs, and the contig assemblies as dark blue glyphs.

Two BAC sequences were then selected for more in-depth analyses as follows.

5.3.4 Selection of BAC clones for gene content: *Serpin* and *rRNA*

In order to select BAC clones for sequence closure, BAC end sequences (BES) [382] were assessed against, the NCBI CDD [148], the BmiGI [152, 361] [383], and the SSH transcripts [151] (Appendix 5.3). The BAC clone BM-005-G14 (GenBank:HM748961) was identified in the BAC end analysis with significant alignment to a *serpin* conserved domain (CDD) [148] cd00172, and BmiGI [152] transcript TC24850. The second BAC BM-012-E08 (GenBank:HM748964) was selected and sequenced based on significant alignment to *Rmi* EST sequence BEAE880F/R, a transcript highly expressed in tick responding to cattle [151].
The following result section describes the genomic; gene and comparative analyses for the BAC sequences BM-005-G14 and BM-012-E08.
The following section 5.3.5 details the analyses for BAC BM-005-G14 a low repetitive, gene rich genomic region.

5.3.5 BAC BM-005-G14 assembly and analysis

The BAC clone BM-005-G14 was sequenced at 6.7x coverage (1,536 Sanger reads, insert size 135Kb). The reads were *de novo* assembled with phred/phrap [364] into six contigs greater than 2Kb and length 136,422Kb. The BES positioning in two contigs confirmed the correct contig assembly. The final contig set was ordered and oriented by read linkage results, BES positioning and gene annotations. The BAC sequence was finished with gap closure into a 135Kb genomic sequence (GenBank:HM748961). Gene prediction and comparative

analysis identified regions of similarity to seven genes displayed in Figure 5.4. The forward strand contained: a *papilin* with a CDS length of 8,361bp consisting of forty exons that span BAC sequence position 2,190 to 88,307bp; a *helicase* with a CDS length of 4,800bp consisting of four exons that span BAC sequence position 6,015 to 14,766 bp; a hypothetical protein (H1) with a CDS length of 2,394bp consisting of eleven exons that span BAC sequence position 93,878 to 10,9076 bp. On the complementary strand; a pogo transposable element with a CDS length of 615bp consisting of three exons that span BAC sequence position 49,728 to 50,977 bp; a hypothetical protein (H2) with a CDS length of 720bp consisting of two exons that span BAC sequence position 110,728 to 111,698 bp; a hypothetical protein with a CDS length of 2,931 bp consisting of eleven exons that span BAC sequence position 112,452 to 122,035 bp. The hypothetical protein was conserved to *Isc* and similar to an endonuclease reverse transcriptase (ERT) in *Bos taurus,* the predicted CDS also contained a *serpin* domain (see later serine protease inhibitor result section). A final *serpin* with a CDS length of 2,766 bp consisting of ten exons that span BAC sequence position 123,297 to 133,688 bp (Figure 5.4).

Two genes of particular interest to the study were the serine protease inhibitor (*serpin)* (cd00172), originally targeted to select this BAC sequence, and the large multiple domain *papilin* gene spanning approximately 90Kbp of the 135Kb BM-005-G14 BAC sequence. The *papilin,* an extracellular matrix glycoprotein that shares a conserved protein domain order in orthologous genes was then selected for further investigation.

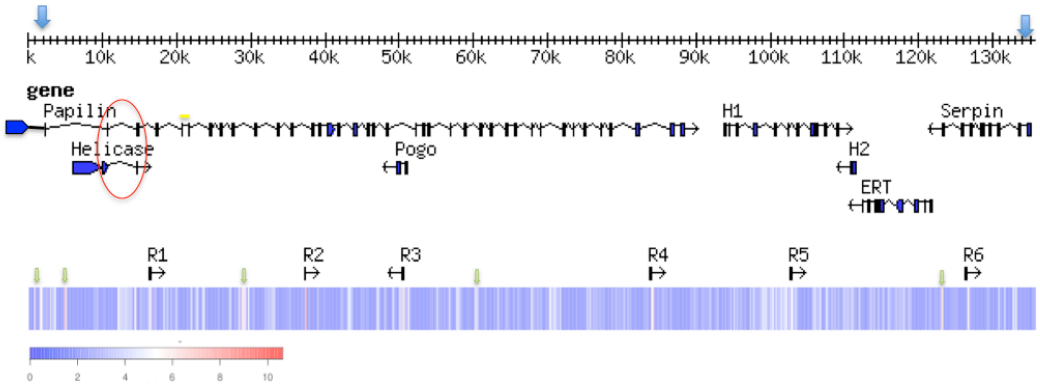


Figure 5.4 BAC BM-004-G14 (135 Kb) gene structure and Cot sequence frequency. Genes and Ruka elements (R1-R6) are displayed over a heatmap of Cot sequence frequency (log2), blue (low) = 0 and red (high) = 10, genes include a *papilin*, a nested helicase-like protein, a pogo element protein, two hypothetical proteins, an endonuclease reverse transcriptase protein and a *serpin*. Low complexity sequence is shown over the heatmap (green arrow). Exon overlaps are circled in read.

5.3.5.1 *Papilin* and *Helicase* cDNA: resolving nested genes

The final *papilin* product of 8,761 bp, was merged from three cloned products, the 5' Race to primer AdamS_R1 product length of 867 bp, primer regions papilin57383F to PapilinR3 product length 7,723 bp and pap12440F to pap13230R direct sequence product length 813 bp (primers can be found in Appendix file 5.2). The conserved domains are as follows: a Thrombospondin type 1 protein (TSP) domain (pfam00090) positioned 349-510 bp, an ADAM-TS Spacer 1 positioned 823-1167 bp (pfam05986), a set of four TSP domains in sequence positions 1204-1371,1387-1548,1561-1737,1724-1896 bp (pfam00090); ten BPT1/Kunitz family domains (KU) (cd00109) serine protease inhibitors can be found at positions 4654-4815,4831-4992, 5008-5169, 5185-5343, 5371-5532, 5545-5706, 5749-5907,5920-6081,6121-6279,6355-6510 bp; a whey acidic protein-type four-disulphide core

domain (WAP) (pfam00095) in position 6901-7065 bp; a set of three immunoglobulin family (IG) (pfam07679) domains in positions 7198-7434, 7447-7680, 7864-8979; and a final protease and lacunin domain (PLAC) (pfam08686) positioned in the 8110-8208 bp region.

Nested in intron 2 of the *papilin,* and on the same coding strand, is the *helicase* gene. This *helicase* gene overlapped exon regions with *papilin* exons 2 and 3 (Figure 5.1). *Helicase* exon 3 position 9,987-10,727 bp and *papilin* exon2 position 10,625-10,727bp share 102 bp. The second shared exon region of 86 bp length was located between *helicase* exon4 position 14680-14766 bp and *papilin* exon3 position 14,680-14,813 bp. The shared overlap regions, circled in red in Figure 5.4, are shown in more detail in the sequence alignment Figure 5.5.

Figure 5.5. Sequence alignment of exon overlap between *papilin* position 502-738bp and *helicase* 3 position 974-4802bp. The consensus black bar indicates the region of overlap. Blue arrows indicate exon junctions, *papilin* T-G positions 604 and 605 and *helicase* G-C positions 4715, 4716.

5.3.5.2 *Papilin* and *Helicase* qRT-PCR: gene expression in tick life stages

The expression of the *papilin* and *helicase* were determined in a number of tick life stages. The gene expression fold change relative to pooled cDNA for a number of life stages were tested for both *papilin* and *helicase* genes. In quantitative real-time PCR (RT-PCR) analysis, it was demonstrated that expression of the *papilin* gene (grey bar) was the strongest in tick larvae sensing and trying to attach to the host (Figure 5.6). The *helicase* (white bar) shows greatest up regulation in the ovaries of female ticks semi-engorged (17days old) attach to the host. The *papilin* (grey bar) also showed differential up regulation in the ovaries. RT-PCR confirmed differential expression between these two genes in at least two tick life stages tested.

As the *papilin* had increased expression in tick larvae sensing and trying to attach to the host, sequence level differences were examined between ixodid (tick) and mammalian (host) species for this gene, which is important in host and tick interaction.

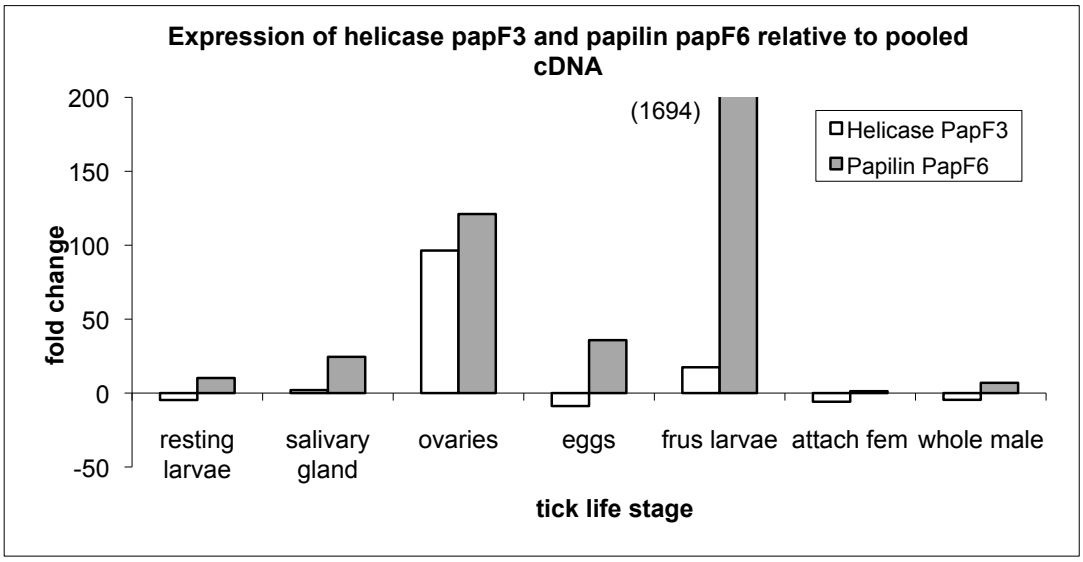


Figure 5.6. Summary of *helicase* and *papilin* qRT-PCR fold change expression in tick life stages relative to eggs, graph bars are *helicase* (white) and *papilin* (grey).

5.3.5.3 Tick *papilin* comparative studies: Identifying tick-specific sequence differences

The mRNA sequence for *C. elegans* (NM_072616.3), *D. melanogaster* (NM_176574.2), *A. mellifera* (XM_396472.3), *I. scapularis* (preliminary data set cDNA jcvi 0.5 set 35859.m000024) and *B. taurus* (XM_002700672.1) was summarised to view domain differences, see Figure 5.7. The domain structure and number was closely conserved in the invertebrate species. Of interest however, the *helicase* domain nested in *Rmi* was also found in the *Isc*. The *Isc papilin* sequence is not found, and it is not clear why this is the case, in the later release 1.1. The number of conserved domains differed the greatest in *Bos taurus* (GenBank: XP_002700718.1) as compared to *Rmi*. The full *papilin* protein multiple sequence alignment between *R. microplus* and *Bos taurus* (XP_002700718.1) can be found

in Appendix file 5.4. These differences in domains include an extra full TSP domain and two fragments highlighted in blue in *Rmi*, a single bovine BTI/Kunitz serine protease inhibitor compared to the set of ten in *Rmi* (red) and the absence of a WAP domain upstream of the IG-set. The multiple protein sequence alignment of *Rmi* 2,180-2,335 bp, human (sp|O95428.4) 999-1,046 bp and bovine 963-1,009 (XP_002700718.1) displays the WAP domain region, boxed in blue in *Rmi* (Figure 5.8) as absent in the mammalian sequences for *H. sapiens* and *B. taurus*.

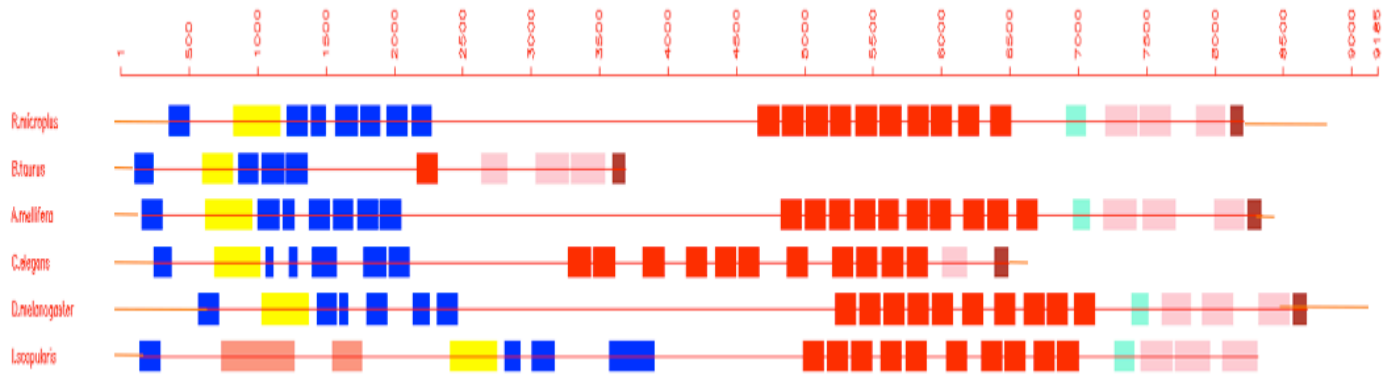


Figure 5.7. Comparative domain analysis between *papilin* orthologues: *Rmi papilin* mRNA (8761 bp); *Bos taurus* (3794 bp); *A. mellifera* (8424 bp); *C. elegans* (6654bp); *D. melangaster* (9171bp); and *I. scapularis* (8328bp). Types and number of domains are displayed, Thrombospondin (blue), Adam-TS (yellow), Kunitz (red), WAP (light green), Ig-set (pink) and PLAC (purple-brown). The helicase domain (brown) in *I. scapularis* is nested between the first Thrombospondin and Adam-TS.

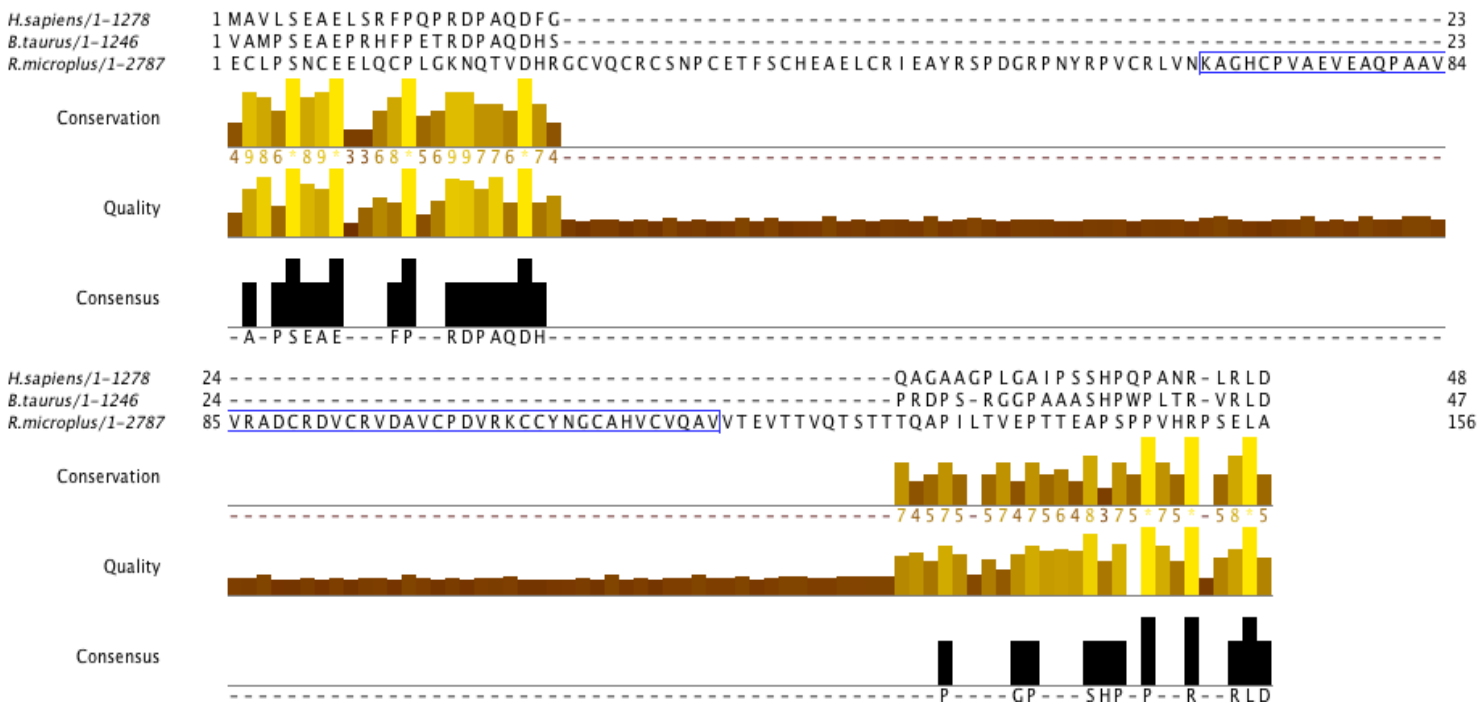
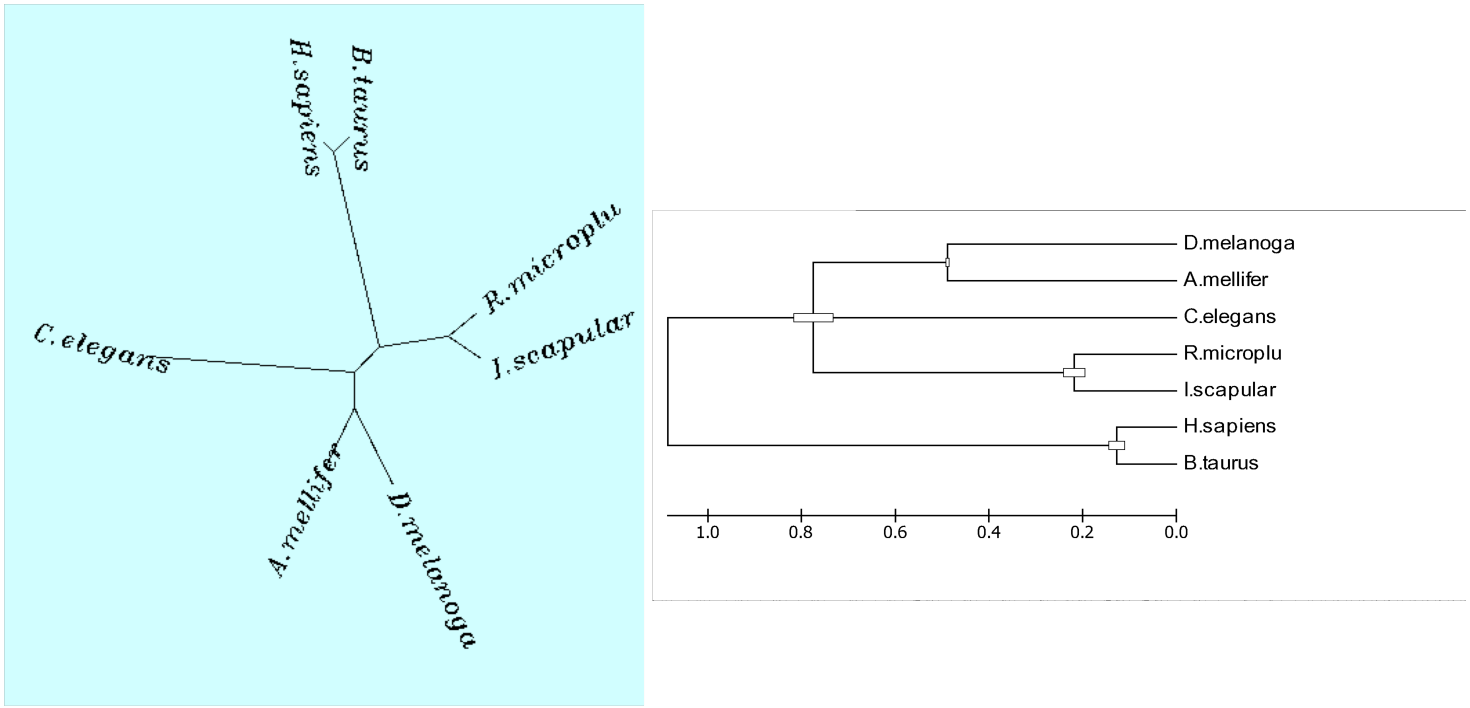


Figure 5.8. Protein sequence alignment of *Papilin* WAP domain, sequence regions *Rmi* 2,180-2,335, *Bos taurus* 963-1,009 and *Homo sapiens* 999-1,046. The *Rmi* WAP domain, position, 2,247-2,297bp, (light blue) in absent in mammalian sequences.

Multiple sequence alignment [34] and phylogeny analysis [375] produced a mammalian clade for *Bos taurus* (XP_002700718.1) and *Homo sapiens* (NP_775733.3) and a tick clade for *Rmi* and *Isc* (35859.m000024_1), a hexapod clade for *D. melanogaster* (NP_788752.2) and *Apis mellifera* (XP_396472.3) and a single node for *C. elegans* (NP_505017.1) as shown in Figure 5.9A. Evolutionary analyses shows that mammalian (host) *papilin* diverge at an earlier time than the divergence of hexapoda *papilin* from tick *papilin* (Figure 5.9B).



A B

Figure 5.9. Phylogenetic analysis of papilin protein for *Bos taurus* (XP_002700718.1), *Homo sapiens* (NP_775733.3), *Rmi*, *Isc* (35859.m000024_1), *D. melanogaster* (NP_788752.2), *Apis mellifera* (XP_396472.3) and *C. elegans* (NP_505017.1). The trees are represented A) Neighbour-joining and B) Molecular clock.

The *serpin* downstream of the *papilin* on the negative strand was next investigated, to identify if the *Rmi papilin-serpin* gene synteny was conserved in other species.

5.3.5.4 Serine protease inhibitor: Serpin pseudogenes

Downstream of the *papilin* gene*,* a full CDS for *serpin* was predicted. The predicted *serpin* domain structure, however, was fragmented with the N-terminus and C-terminus rearranged, exon2 residues 266-364 and exon9 1-63 residues. Attempts to sequence the *serpin* cDNA resulted in a 500bp product. A single PCR product based on forward primer (SerpF3) in exon5 and reverse primer (SerpR2) in exon9 was sequenced (Appendix 5.5). The small product sequenced matched only 148 bases of predicted exon5 and 231 bases of predicted exon9 (exons 6-8 were not in alignment). Conserved *serpin* domain analysis found also two fragments in predicted ERT gene exon2 residues 115-189 and exon4 residues 185-364. To determine whether the adjacent position of a *serpin* with *papilin* is common, a search of mosquito, fly genomes and *Isc* found no evidence of a *serpin* downstream from the *papilin* indicating this arrangement as not conserved within arthropods.

5.3.5.5 BAC and Cot comparison: element genome wide frequency estimation

To gain better insight to genomic structures a Cot DNA comparison to the BAC sequence was undertaken. DNA reassociating kinetics based Cot filtration of genomic DNA was used to reduce the concentration of repetitive DNA sequences that dominate the *Rmi* genome, in order to analyse the "gene-rich" single/low-copy and the moderate repetitive DNA fractions [362, 384]. Two fractions of moderate

to low repetitive regions of *Rmi's* genome were selectively obtained from Cot filtration [362] and then analysed with BM-005-G14 BAC sequence, to assay the frequency of specific BAC sequence within the entire genome. Cot696 and Cot69 DNA 454 sequences were mapped to BM-005-G14 to determine the frequency that the BM-005-G14 sequences were found in the Cot selected fraction of the tick genome. The read depth in a 100bp window over total mapped read (bp) was calculated and the log2 value plotted as a heatmap (Figure 5.4). Six SINE *Ruka* elements [355] (R1-R6) were identified in BM-005-G14 (Figure 5.4) at positions 16,283-16,459, 37,204-37,398, 50,357-50,535, 83,819-83,996, 102,593-102,771, 126,277-126,452, the Pogo gene appear in the white-red bands, and (as expected) low frequency *papilin* and *serpin* in blue bands. Other regions of high frequency were identified as a 321bp ATCT repeat positioned at 4,864-5186 and, a 124bp TTTC repeat positioned at 878-1,004bp and 241bp CAA repeat in region 122,719-122,961, these are shown as green arrows over the heatmap in Figure 4.4. The overall estimate of BAC sequence coverage was 38.80% and 41.59% respectively for Cot 696 and Cot 69.

Based on the proportion of mapped reads relative to the total sequenced reads from the two Cot DNA experiments the genome wide frequency of a single 195bp *Ruka* element (R2) at position 37,204-37,398 was estimated based on the extrapolation of the two Cot fractions back to time zero to represent 0.42% (29Mb) of the 7.2Gb genome (Appendix 5.6). Although this estimate is approximate, the frequency of this specific *Rmi* Ruka element in the genome is estimated to be at

least 152,923 copies.

5.3.6 BAC BM-012-E08 assembly and analysis

The BAC clone BM-012-E08 was sequenced (1536 Sanger reads) at an expected size based on restriction digest of 65Kb (not shown). Due to the complexity of BM-012-E08 BAC, the assembly metrics were tested to de-convolute the sequences (Appendix 5.7). In summary, a conservative assembly approach [165] assembled less reads and produced more singletons but increased the length of the total assembly. BM-012-E08 was assembled into a 52Kb consensus based on 18 contiguous sequences greater than 1kb in size.

As very few arthropod retroelements and a large percentage of small RNA (17%) were identified with RepeatMasker [380], eight *de novo* interspersed repeat motifs were identified and masked for gene predictions. An almost complete, 18S ribosomal RNA gene, internal transcribed spacer 1 (ITS), 5.8S ribosomal RNA gene, ITS 2, and 28S ribosomal RNA gene, was identified by best sequence similarity to *Amblyomma americanum* (GenBank AF291874) (Appendix 5.8A). In the assembly, there is evidence of at least three ribosomal 18S 5' and 28S 3' units (Appendix 5.8B). In Figure 5.10, the rDNA BAC positions are shown as blue arrows, BAC end positions in green, and the intergenic regions in brown. The remaining sequence not accounted for contained repetitive DNA sequence, similar to the highly repetitive intergenic regions found between the repeating rDNA units. In Figure 5.8, the dot matrix of the best (parsimonious) ordered and oriented

contigs display three rDNA units and repetitive intergenic regions aligned against a single unit.

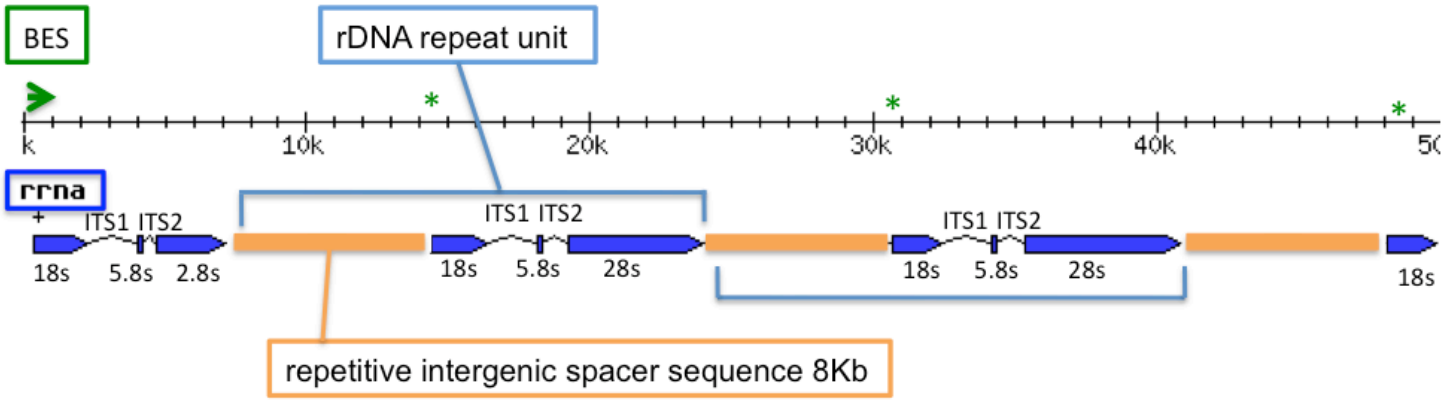


Figure 5.10 BAC BM-012-E08 50Kb sequence, with a repetitive 8Kb intergenic spacer (brown), and rRNA 18S, ITS1, 5.8S, ITS2 and 28S (blue) together make the repeat unit structure (light blue). The BES start position are indicated by a green arrow and *copies.

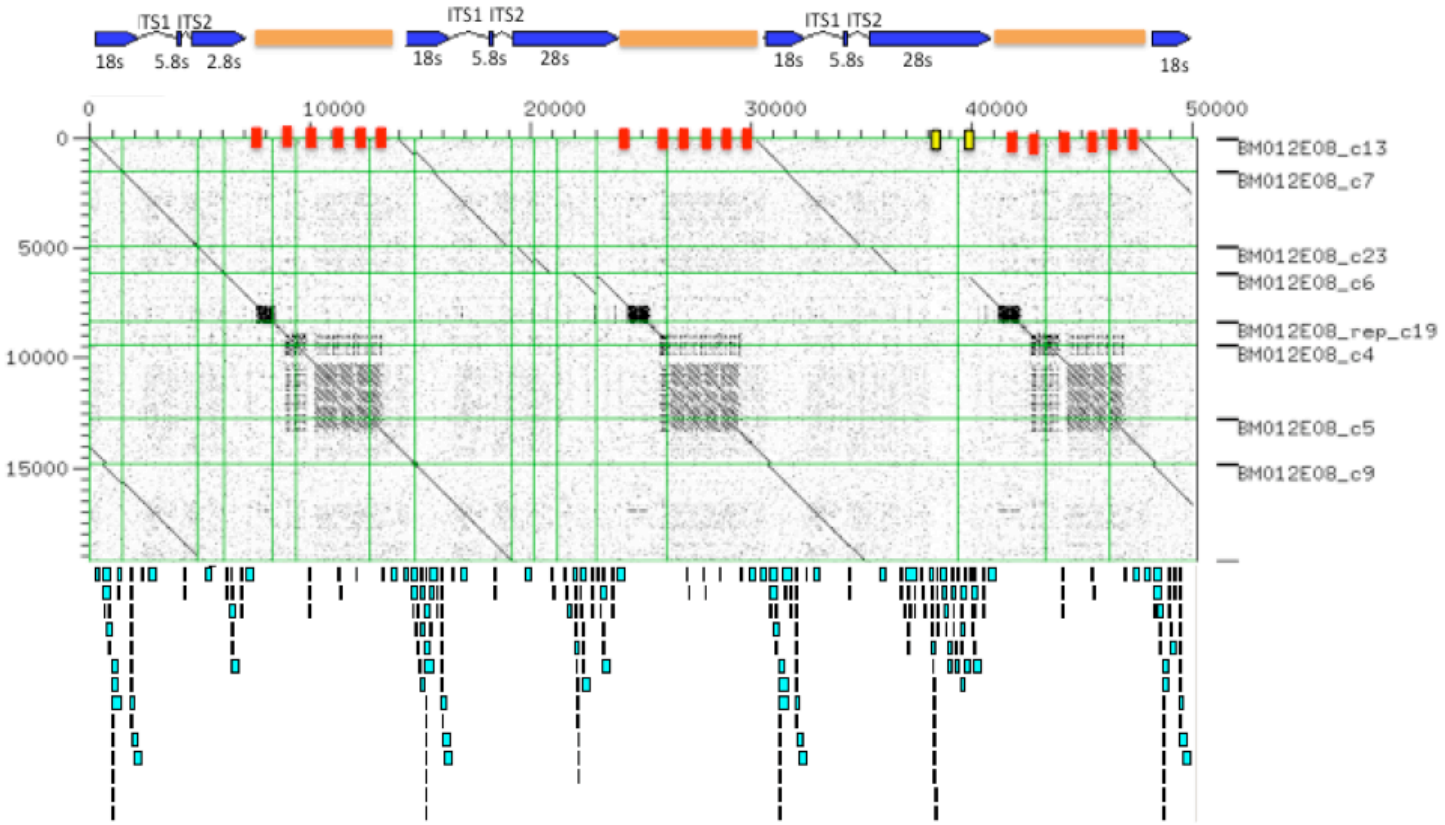


Figure 5.11 Sequence dot matrix of BAC BM-012-E08 displays highly repetitive regions and rRNA (blue) fragments. Horizontal axis shows regions confirmed (red) by PCR and not confirmed (yellow). The mapped Cot696 454 read are displayed below the dot matrix.

The junction of the repeat motifs and the rRNA were tested by PCR, direct sequenced and clones sequenced to validate sequence and assembly accuracy. Three BAC repeat junction positions were confirmed by PCR and shown in Figure 5.11 as red markers on the horizontal axis of the dot plot (17K, 22K and 38K). Primer sequences and gel pictures can be found in Appendix 5.2 and Appendix 5.9 respectively. The marker 22K downstream of the 28S unit in the large repeat region (dark blue glyph) was confirmed by PCR, (lanes 1 and 2, Appendix 5.9A). The marker 38K upstream of the 18S unit in the large repeat region was also confirmed by PCR (Lanes 5 and 6 Appendix 5.9A). Marker 28K confirmed the region downstream of the 28S unit in the smaller dense repeat region in contig6, shown in Figure 5.11, this is repeated in contg-rep_c50 and contig1.

5.3.6.1 Ribosomal DNA (rDNA) structure analysis

At least three full ribosomal 18S and 5.8S repeat units could be ordered although the 28S unit assembly was partial in the first repeating unit1, and in unit 3 contained a break point ITS (1Kb) as compared to the *Ambylomma americanum* rRNA sequence. The rDNA repeat elements and ribosomal units were then tested by PCR to validate the gene size and to reveal more detail on the rDNA repeat unit and a putative interrupter sequence in LSU sequence that had previously been identified in the *Drosophila* genome [385]. Long range PCR confirmed rRNA unit size of 7-8Kb (lane 2) and a large intergenic repetitive region of 8-9Kb (lane 6) (Appendix 5.9B). The 28S breakpoint/ large interrupter region could not be confirmed by PCR, position is highlighted yellow in Figure 5.11 (~42Kb), tested

primer sets 15K and 18K can be found in Appendix 5.2. The BAC assembly and long range PCR confirmed the rRNA unit size of 8Kb and a large intergenic repetitive region of at least 9Kb. The *Rmi* 18S sequences were then analysed for tick specific differences.

5.3.6.2 Tick ribosomal DNA comparative studies: Identifying tick-specific sequence differences

To identify sequence differences in rRNA protein binding sites the 18S small subunit (SSU) was aligned to the 16S *E. coli* SSU and bovine18S.

Three *E. coli* complex binding sites [386] were analysed and a number of sites with conserved changes in tick as compared to mammals were found. Sites were also identified conserved in *E. coli* and mammals, and a conserved change found in tick (Figure 5.12).

In Figure 5.12, regions for rRNA protein binding [387] have been highlighted to identify regions of differences in *E. coli* 16S (Embl: X80725) and tick and mammalian 18S sequences. Represented in the multiple sequence alignment are five tick species, *A. americanum* (GenBank:AF291874.1*), A. glauerti* (GenBank:AF115372.1)*, A. variegatum* (GenBank:L76346.1)*, A. tuberculatum* (GenBank:L76345.1)*, A. maculatum* (GenBank:L76344.1) and *R. microplus* 18S RNAmmer prediction [388], the tick host *B. taurus* (GenBank:DQ222453.1) and *H. sapiens* (GenBank:NR_003286.2). Protein binding sites are highlighted in the following *E. coli* sequence regions: S7_S9_S19 complex (red) 936-965, 972-991, 1208-1262, and 1285-1379bp; S8_S15_S17 complex (orange) 118-240, 576-606, 629-685 and 706-769; S8_S17 complex (green) 1406-1442 and 1462-1494bp

[386]. The multiple sequence alignment and annotation can be found in Appendix 5.10. The protein rRNA binding site for protein S7 in *E. coli* (1236-1240, 1373-1383bp) shows a single SNP between the tick and mammalian sequences (Appendix 5.10). In Figure 5.9 an overview of Appendix 5.10 alignment is shown centre, the alignment subset (bottom blue box) shows a highly conserved region between 18S and 16S sequences (*R. microplus* 1122-1160bp) with an example conserved SNP change G->A in tick, this was confirmed in all three 18S units. The alignment subset (top box) (*R. microplus* 791-818bp) shows more sequence differences for the *E. coli* S8_S15_S17 binding complex region. In the *Rmi* 18S sequence, a total of 63 *Rmi* nucleotide positions showed conservation in all tick species and a conserved change in *B. taurus* and *H. sapiens*.

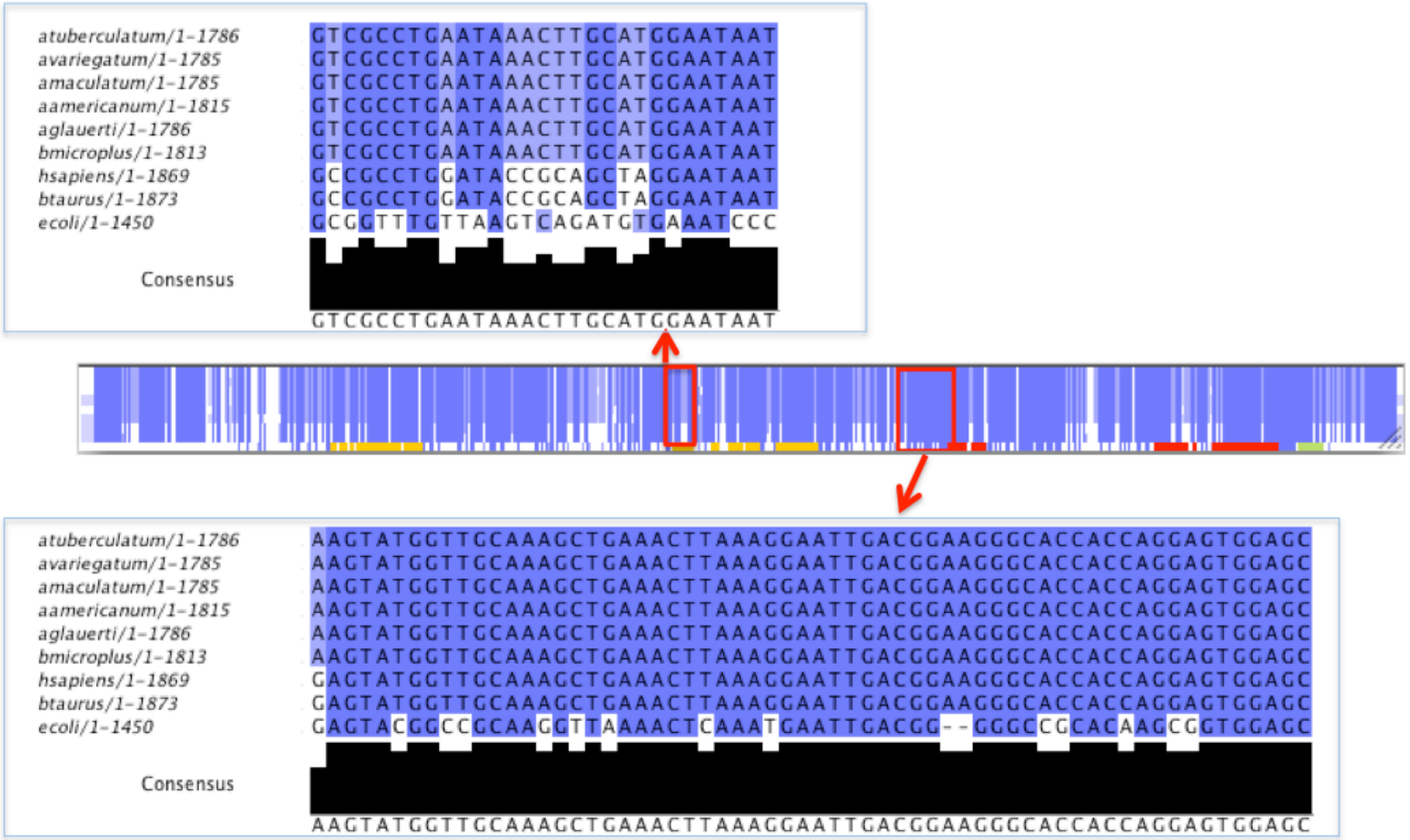

atuberculatum/1-1786 GTCGCCTGAATAAACTTGCATGGAATAAT
avariegatum/1-1785 GTCGCCTGAATAAACTTGCATGGAATAAT
amaculatum/1-1785 GTCGCCTGAATAAACTTGCATGGAATAAT
aamericanum/1-1815 GTCGCCTGAATAAACTTGCATGGAATAAT
aglauerti/1-1786 GTCGCCTGAATAAACTTGCATGGAATAAT
bmicroplus/1-1813 GTCGCCTGAATAAACTTGCATGGAATAAT
hsapiens/1-1869 GCCGCCTGGATACCGCAGCTAGGAATAAT
btaurus/1-1873 GCCGCCTGGATACCGCAGCTAGGAATAAT
ecoli/1-1450 GCGGTTTGTTAAGTCAGATGTGAAATCCC
Consensus
GTCGCCTGAATAAACTTGCATGGAATAAT
atuberculatum/1-1786 AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC
avariegatum/1-1785 AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC
amaculatum/1-1785 AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC
aamericanum/1-1815 AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC
aglauerti/1-1786 AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC
bmicroplus/1-1813 AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC
hsapiens/1-1869 GAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC
btaurus/1-1873 GAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC
ecoli/1-1450 GAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGG--GGGCCGCACAAGCGGTGGAGC
Consensus
AAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGC

Figure 5.12. Overview of multiple sequence alignment; six tick species 18S *A. americanum, A. glauerti, A. variegatum, A. tuberculatum, A. maculatum* and *R. microplus*; two mammalian species 18S tick host *B. taurus, H. sapiens,* and *E. coli* 16S. *E. coli* 16S protein binding sites are highlighted S7_S9_S19 complex (red), S8_S15_S17 complex (orange), S8_S17 complex (green) (Weiner et al. 1988). Top box shows multiple sequence alignment of binding site sequence variance (*Rmi* 791-818bp). Bottom boxed multiple sequence alignment shows tick / mammalian conserved SNP *Rmi* 1122-1160bp

To examine if this BAC sequence had the same variation at the R2 retrotransposon target site of the LSU [389], the *Rmi* rDNA LSU positioned at 19240-23971 bp (19K unit) and 35316-40798 bp (35K unit) were aligned. The specific variation of conserved SNPs guanine (G) and thymine (T) previously found for *Ixodidae* (hardback ticks) were confirmed, at positions 2,800 bp (G) and 2,801 bp (T) for the 19K unit and 2,873 bp (G) and 2,874 bp (T) for the 35K unit. No R2 retroelements were identified, a small fragment of LINE/R1 element TRAS9_SC was however found in the 35K unit at position 1,801-1,889 bp.

5.3.6.3 BAC and Cot comparison: element genome wide frequency estimation

To estimate the genome wide frequency of elements in this highly repetitive BAC, a single Cot fraction Cot 696 was used to align to BM-012-E08. The reads were mapped to the BAC sequence at 100% identity and 90% read coverage, shown beneath the dot matrix in Figure 5.11. Reads could not be aligned to the densely repetitive rDNA intergenic regions, it was also noted that the short interspersed Ruka element found in BM-005-G14 was absent in BM-012-E08.

5.3.7 Bioinformatics workflow for vaccine candidate identification

Effective and automated analysis of transcriptome sequences is critical with the advent of next-generation sequencing. This is even more of a challenge when dealing with largely uncharacterized tick transcriptome, and without a reference genome. In the absence of high throughput (HTP) bioinformatics pipelines for vaccine candidate identification (VI), a novel pipeline was constructed to identify vaccine candidates for bovine tick resistance using a global transcriptome approach. Figure 5.13 shows the bioinformatics workflow steps to prepare the data for high through put analysis for vaccine identification.

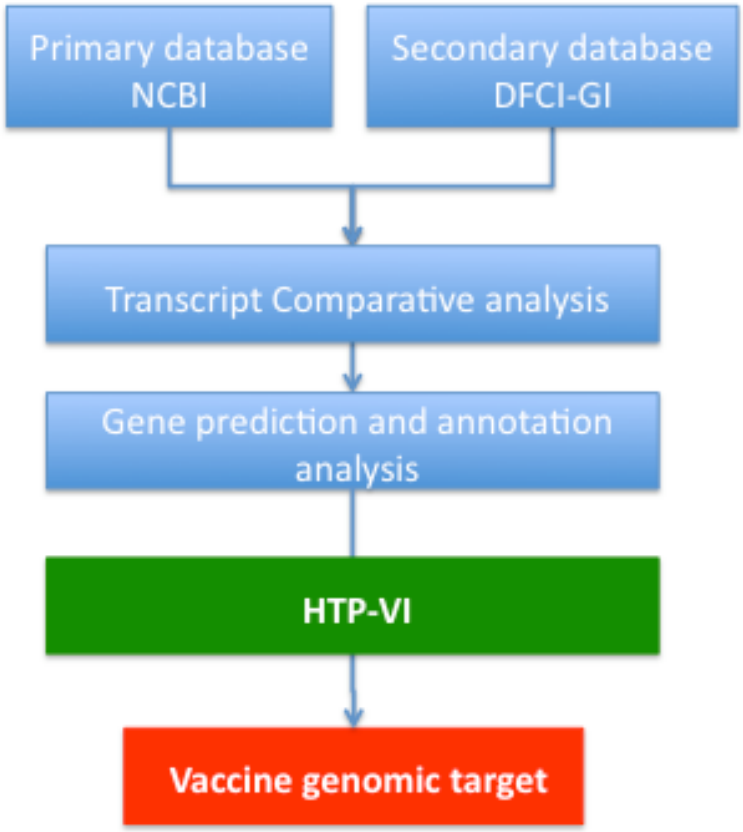


Figure 5.13 Workflow steps to prepare data for peptide analysis. The blue-boxed steps prepare the data for the green-boxed HTP-VI bio workflow to identify the vaccine targets (red box).

HTP-VI pipeline draws together many bioinformatics processes for the fast and effective screening of sequences (including unknowns) based on a selection of protein properties to identify candidate genes for vaccine development (Figure 5.14).

To identify vaccine candidates for susceptible cattle, the complete DFCI *Bmi*GI dataset of expressed genes for tick, and selected BAC genes from BM-005-G14

were screened. The database was annotated for domain & protein family identification, cell localization and host homology to ensure that an autoimmune response in the host would not occur. Candidate selection was based on protein localization (surface exposure), function (those functions being important to maintain pathogen feeding on the host), the availability of peptide epitope regions (correct residue physicochemical properties), no homology to bovine, and the unknowns that met criteria other than function. Selected candidates were clustered for B-Cell linear epitope predictions based on conserved domains in ‘knowns’ and predicted ORFs in ‘unknowns’.

In over 14,600 sequences, HTP-VI identified 250+ vaccine candidate sequences that clustered and grouped into 41 known protein families and 58 unknown. To test the surface exposure of predicted epitope peptides, a large protein family with a highly conserved domain, the Histamine Binding Proteins (HBP), was selected for protein 3-D modeling. Further detail can be found in 5.3.8.

HTP-VI effectively identified candidate genes in an under characterized genome for vaccine development. This is the first bioinformatics approached developed of this kind and on this scale. The identified pathogen peptides have been validated in host sera, and have entered vaccine trials.

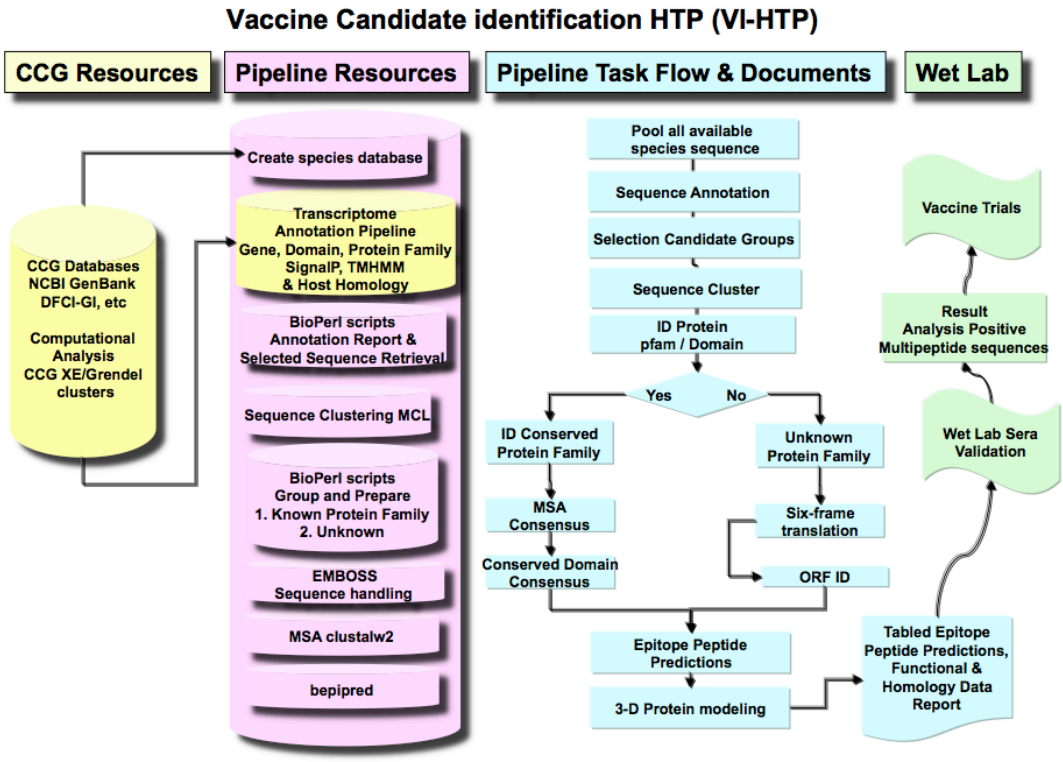


Figure 5.14 High ThroughPut Vaccine Identification bioinformatics pipeline, HTP-VI.

5.3.8 Vaccine candidate tests

Out of the 250+ candidates identified, based on qPCR and microarray analysis a total of 81 candidates were selected. A total number of 1800+ epitope regions with a peptide length greater than or equal to 8 residues were predicted for the BAC genes and BmiGI, conserved subsets were then screened in sera and 14 candidates selected for animal trial [381]. Focusing on 48 BAC peptides tested in sera, the following three genes *papilin*, *serpin* and *tetraspanin* tested positive. In Table 5.1, 6 tested positive in susceptible sera only, 2 tested positive in resistant sera only, and 8 tested positive in both susceptible and resistant sera.

A small number of candidates were supported by proteomic analysis conducted on unfed larvae [390] and microarray analysis in an acaracide-inducible gene expression experiment [339].

Table 5.1 IgG Bovine sera B-Cell peptides results for BAC genes *papilin, serpin* and *tetraspanin*

Label	**Peptide**	**Sera result**
Papilin1	GCCPDGETPAEGPDNE	Positive susceptible
Papilin 2	GCCPDGRTPARGPEYDG	Positive resistant
Papilin 3	CHGNNNRF	Positive resistant and susceptible
Papilin 4	GGCQGNANNFATEDECS	Positive resistant and susceptible
Papilin 5	GCDGNENN	Positive resistant and susceptible
Papilin 6	GGCEGNDNRF	Positive resistant
Papilin 7	CGGNRNRF	Positive susceptible
Papilin 8	ELNCKPRG	Positive susceptible
Papilin 9	AMSDGGEYSCQADNGHSNQS	Positive resistant and susceptible
Serpin bac	RWNTPFDPCRTVPGDFH	Positive resistant and susceptible
Tetraspanin_bac_139_189 Tetraspanin 3.2	TSTPPNRSAYTTENREE	Positive resistant and susceptible
Tetraspanin 3.3	YTTENREEGGEHGPGVP	Positive susceptible
Tetraspanin 3.4	GEHGPGVPPPGATPLMR	Positive susceptible
Tetraspanin 3.5	PGATPLMRNETKTPEGE	Positive susceptible
Tetraspanin_bac_28_38 Tetraspanin 2.1	VPAFQEDEGAV	Positive resistant and susceptible
Tetraspanin_bac_45_53 Tetraspanin 3.1	RDKRPKEHV	Positive resistant and susceptible

In the HBP family cluster of 48 sequences, the lipocalin sub families are represented. The phylogenetic tree in Figure 15.15 shows the clades for lipocalin and HBP sequences. Based on a single clade HBP sequences were selected for high conservation and epitope properties (Figure 5.16). The epitope prediction for the histamine binding protein family cluster of conserved sequence was then shown in a 3-D protein structure.

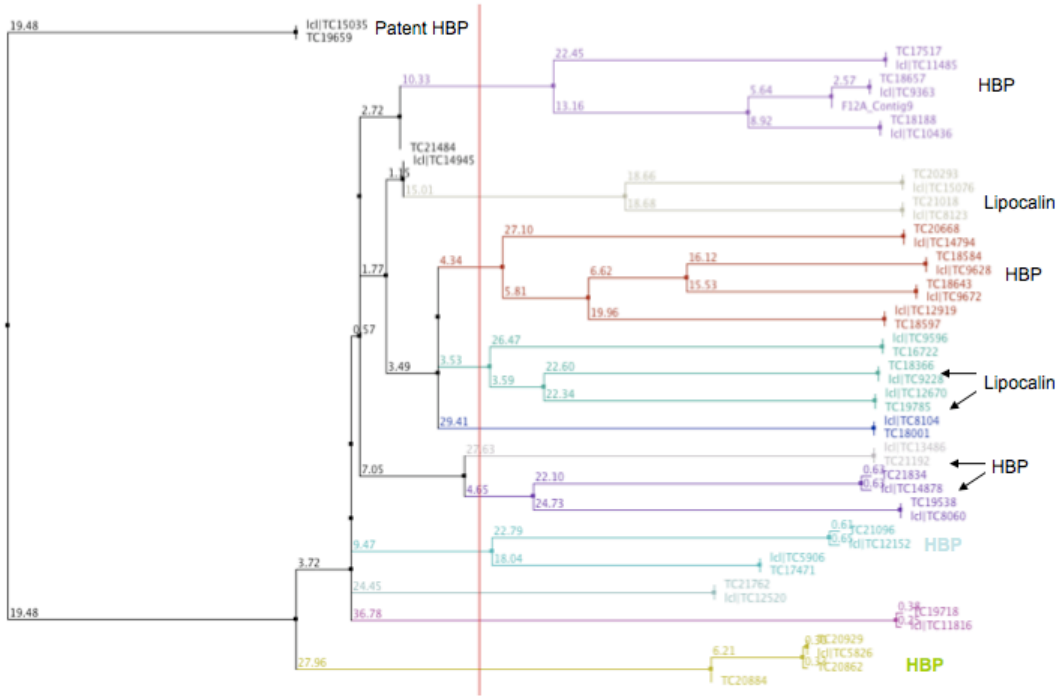


Figure 5.15 Phylogeny of Histamine Binding Protein (HBP) family for epitope conservation analysis, the red vertical line defines the clade colours

CLUSTAL W (1.83) multiple sequence alignment

```
TC18188     ----------PVWADEAANGAHQDALKHLK
TC10436     ----------PVWADEAANGAHQDALKHLK
TC9363      QVKGNKPVWADEAANGAHQD------
TC18657     QVKGNKPVWADEAANGAHQD------
TC17517     E I RADKPAWADEAANGAHQD------
TC11485     E I RADKPAWADEAANGAHQD------
                 * .**********
```

Figure 5.16 Histamine binding protein conserved epitope prediction

The *Rmi* 3-D sequence structure was modelled on the *R. appendiculatus* "*Rhiap* PROTEIN (FEMALE-SPECIFIC HISTAMINE BINDING PROTEIN) lipocalin" which has two chains 1QFT.pdb chains A and B. The structure 3-D modeller template

was based on 1QFT and 1QFV and query TC17517 (Figure 5.17) using the iMOL application [391]. Figure 5.17 (A) shows the 3-D protein structure for female histamine binding protein 2 and the predicted HBP family conserved epitope peptide domain (blue in alpha chain and red in beta chain) that are surface exposed in *Rmi.* Figure 5.17 (B) shows the 3-D HBP structure in *R. appendiculatus* (left) and *R. microplus* (right) with the conserved epitope regions in sphere atom style.

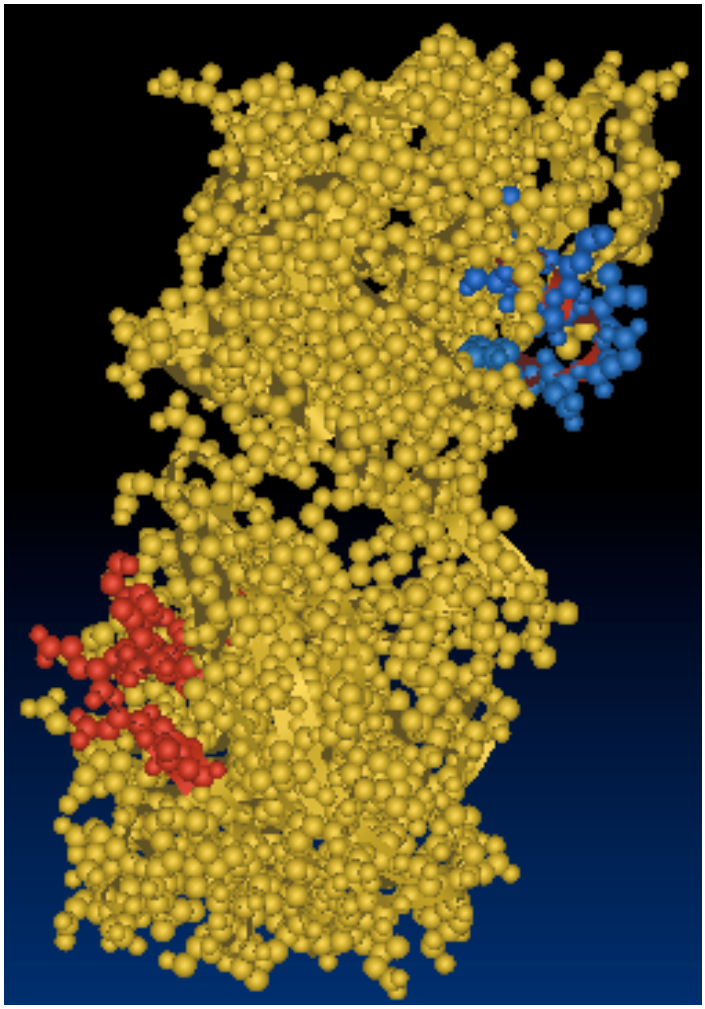


Figure 5.17 (A) 3-D Histamine binding protein surface-exposed conserved epitope shown as blue in alpha chain and red in beta chain

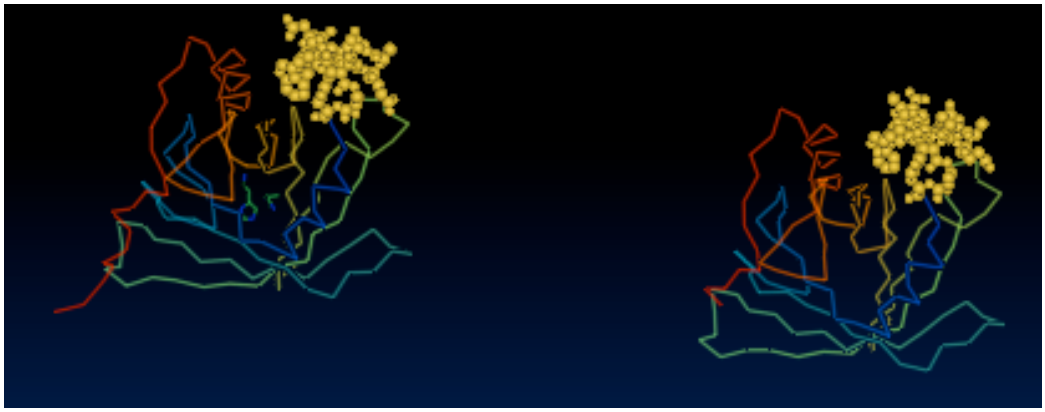


Figure 5.17 (B) *R. appendiculatus* (left) and *R. microplus* (right) 3-D HBP with conserved epitope regions in yellow sphere atom style.

5.4 Discussion

In the absence of a reference genome we describe the *de novo* assembly and in-depth analysis of two *Rmi* BAC clones, selected for gene content important for tick growth and development. In this study of two very different BAC regions newly reported features for eukaryotes and *Chelicerate* genomes are described. The following discussion sections address the study of both the selected BAC sequences.

5.4.1 Tick genomic structure: assembly and predictive models

The correct assembly of the tick genome is a challenge due its repetitive nature, and the lack of predictive models for gene structures. The Cot DNA assembled with 70% of the BES, this confirmed the composition of Cot selected DNA for the moderate repetitive and unique portions of the genome. The BAC assembly, due to low genome level synteny with the *Ixodes scapularis* assembly, depended on the comparative analysis of transcript and the positioning of BES.
The correct assembly of the BM-005-G14 contig set was dependant on the correct BES positioning, while the ordering and orientation was aided by the BES positioning and transcript alignments.

Given the complexity of BM-012-E08, different assembly tools were trialled under different options. Assembly tools with uniform read distribution take a cautionary approach in contig building, and sometimes create two contigs when it could have

created one. This feature reduces over-compression of repeats during the contig building phase and ensures that, for example rRNA stretches which are present multiple times in a the genome will also be present approximately the same number of times in the result files [165]. The repeat motifs, intergenic regions, rRNA sequences and rDNA unit size were confirmed by PCR. However the assembly gaps and insertions show clear deviations from the perfect repeat unit size. This is the first *Rhipicephalinae* assembly of rDNA and the first known attempt at assembly in *Arthopoda* of three external intergenic repetitive units between the rDNA repeating subunits.

5.4.2 Tick gene structure: predictive models

The complexities of gene predictions included intronic regions of nested repeat elements, multiple short exons and overlapping regions complicating the delineation of exon coding regions. Overlapping genes have been reported in *Drosophila* but these genes were on different strands [392]. In eukaryotic research this is the first description of same strand gene overlap between two genes, *papilin* and *helicase*.

Papilin is an extracellular matrix glycoprotein. In *Drosophila, papilin* is found to be involved in, (1) forming thin matrix layers during gastrulation, (2) matrix associated with wandering, phagocytic hemocytes, (3) basement membranes and (4) space-filling matrix during *Drosophila* development [393]. The gene is also essential for

normal embryonic development in *Caenorhabditis elegans* [394]. It has been reported that inhibiting *papilin* synthesis in *Drosophila* or *Caenorhabditis* causes defective cell arrangements and embryonic death. Ectopic expression of *papilin* in *Drosophila* causes lethal abnormalities in muscle, Malpighian tubule and trachea formation. It has been suggested that *papilin* influences cell rearrangements and may modulate metalloproteinases during organogenesis [393-396].

This is also the first *Chelicerate* full-length *papilin* cDNA sequence produced. Our *papilin* gene model (refer to methods) was confirmed by other arthropod species. The *papilin*-nested *helicase* was also found within the *Isc* genome supercontigs (version 1) and this inclusion of the *helicase* shows a level of gene synteny is present in this region between the two distant hard tick species.

The function / activities of *papilin* relate to the following specific domains. An interesting gene domain complex the tick derived Kunitz type inhibitors act as antihemostatic factors [397]. Hematophagous organisms must overcome host hemostasis in order to locate blood and maintain its flow during ingestion [398]. Salivary components produce antihemostatic, anti-inflammatory, and immunosuppressive effects that may facilitate feeding, as well as transmission of tick-borne pathogens [398]. The number of *Rmi* KU domains (x10) present compared to bovine (x1) indicates, based on this domains function, an important change in this genes structure for tick survival.

The whey acidic protein-type four-disulphide core domain (WAP) has protein family members that include the whey acidic protein, elafin (elastase-specific inhibitor) known to have anti-microbial activity [399], catrin-like protein (a calcium transport inhibitor), and other extracellular proteinase inhibitors. A significant sequence variance in bovine was the absence of the WAP domain (Figure 5.8).
Isoforms of *papilin* have been found in a number of *Arthropoda* species, six in *Drosophila* and two in *Apis.* Given the size and complexity of the *Rmi papilin*, isoforms may exist that are yet to be investigated.

Helicase was identified nested as a separate gene between the first 5' thrombospondin and the Adam-TS spacer of the *Rmi papilin*. RACE sequencing from the Adam-TS spacer domain exon in the 5'end direction produced the complete *papilin* product minus the *helicase* insertion, confirming our gene model.

The discovery of shared exon regions for 2 eukaryote genes, *papilin* and *helicase*, is quite novel. Nested genes do occur in eukaryotes [392], nested genes in *D. melanogaster* and *C. elegans* have been found exclusively as embedded sequences in introns. It has been reviewed that in *D. melanogaster* nested intronic genes constitutes approximately 6% of the organism's total gene complement, and 85% of these nested genes are predicted to encode protein [392]. For example at the *gart* locus, the *Pcp* gene is nested in intron 1 of the *ade3* gene on the complement strand. A nested ketoreductase was identified in an *A. aegypti* papilin – however not with exon overlap as shown for the *helicase* identified here. In the

mouse genome, 28 overlapping gene pairs had partial overlapping exons, and did not encompass the entire coding sequence of either gene. In the human genome 51 exon overlaps on opposite strands, again were partial. Neither the human nor the mouse genome contains any overlapping genes that share coding sequences on the same strand. Further the majority of nested intronic genes are functionally unrelated and typically not co-expressed with their external host genes. Therefore further functional analysis of this gene's novel arrangement warrants investigation. No *helicase* element was found nested in the bovine intron region of *papilin*.

The initial identification of the *serpin* domain led to the adjacent *papilin* gene described above. No syntenic evidence was found for the down stream serpin region in *Isc*. Full investigation of this gene family within *Rmi* genomic sequence and the *I. scapularis* genome remains.

The genes for ribosomal DNA are tandem repeated clusters in the heterochromatic regions of metazoan genomes [385, 400], in *Drosophila* 77% of heterochromatin sequence is composed of fragmented and nested transposable elements and other repeated DNAs [400]. It has been reported in vertebrates that ribosomal RNA splicing of occurs, and during processing, in mammalian nuclear 28S pre-rRNA, tissue-specific elimination of an intron bearing a hidden break site occurs [401, 402]. An almost complete, 18S ribosomal RNA gene, internal transcribed spacer (ITS) 1, 5.8S ribosomal RNA gene, ITS 2, and 28S ribosomal RNA gene, was identified by sequence similarity to *Amblyomma americanum* (GenBank AF291874;

[389]), which is the only other tick rDNA sequence analysed at this level of coverage. A similar unit was not identified in the *Ixodes scapularis* genome highlighting the difficulty in the assembly of this region. The *Rmi* rDNA units have been identified as novel due to the lack of the R2 retroelement previously identified in other hard tick sequence, even though the R2 retroelement binding site was conserved. The fragmented nature of the LSU makes it possible that the BM-012-E08 BAC clones are derived from the end of an array of rDNA units in the genome where incomplete and rearranged rDNA units may occur.

5.4.3 Tick DNA comparative studies: Identifying tick-specific sequence differences

A number of conserved changes within rRNA protein binding sites between the ticks as compared to mammals were found. The hard tick specific sequence differences (SNPs) were also found in the LSU R2 retroelement target site.
Due to the uniqueness of tick rDNA sequences it is feasible that tick rRNA could be a target for drug development analogous to the use of bacterial rRNA as antibiotic targets. Consistent with this possibility, a number of conserved changes within rRNA protein binding sites between ticks as compared to mammal were found. The hard tick specific sequence differences (SNPs) were also found in the LSU R2 retroelement target site.

5.4.4 Tick gene expression analysis

The qRT-PCR analysis of *papilin* described confirmed differential increased expression in two life stages, most prominently in larvae trying to attach to the host (Figure 5.6). It was also demonstrated that the *helicase* is strongly expressed in ovaries of feeding females. *Helicases* are often utilized to separate strands of a DNA double helix or a self-annealed RNA molecule using the energy from ATP hydrolysis, a process characterized by the breaking of hydrogen bonds between annealed nucleotide bases. The differential expression of both the *papilin* and *helicase* in adult female ovaries suggests that is perhaps a conserved functional arrangement.

The abundance of RNA as identified in the subtraction library study was not surprising due to increased protein production during feeding. It has been reported that the abundance of ribosomal protein coding genes is not unusual for a transcriptome analysis and illustrates the high degree of redundancy found in such libraries, especially the occurrence of numerous sequences coding for proteins involved in protein synthesis such as ribosomal RNA, e.g. 40S, 60S and other ribosomal genes [403].

5.4.5 The analysis of genome sequence via BAC end sequencing and Cot DNA

The *Rmi* genomic DNA that was enriched for single/low-copy and moderately repetitive DNAs [362] along with BAC end sequencing have provided valuable insights into *Rmi* genomic structure. Mapping the reads of the Cot DNA to the BAC sequencing identified regions of high repetitive content in BM-012-E08 complex

intergenic region by the absence of mapped reads. Also moderately repetitive regions could be identified such as the RUKA element in BM-005-G14.
In particular using the two Cot filtrations, we were able to estimate the frequency of any specific genomic sequence within the entire genome. As an example presented, major frequency peaks were identifiable and the relative frequency of the SINE Ruka [355] element in the genome was estimated for BM-005-G14. Although absent from the euchromatic section of the relatively compact 1.8 Gb genome of *Drosophila melanogaster* [404], several distinct families of SINEs with copy numbers of up to 590 K per genome have been described in *Aedes aeygpti*, the mosquito vector of the yellow fever virus [405, 406]
The frequency of *Rmi* Ruka element in the genome was estimated based on a single occurrence on the extrapolation of the two Cot fractions to represent 0.42% of the genome, at least 152,923 copies.

A previous examination of 3 BAC sequences, and the (DFCI) Gene Indices [45] for the four ixodid tick species, *A. variegatum* [407], *R. appendiculatus* (*Rap*) [408], *B. (R.) microplus* [152] and *I. scapularis* [369], estimated that the Ruka repeat sequences comprise approximately 1.6% (4kb) of the 250Kb of *Rap* genome (BAC sequence). Then on the following assumptions (1) that these *Rap* BAC contigs are representative, and (2) a genome size for *R. appendiculatus* of 1Gb, that a total of 65,000 copies of Ruka could be predicted [355]. Since our estimation in *Rmi* is based on a single element we expect the number of Ruka families will occupy a much larger fraction of the genome than previously estimated.

5.4.6 Vaccine candidate identification

The HTP-VI bioinformatics workflow effectively identified candidate genes in an under characterized genome for vaccine development. This is the first bioinformatics approached developed of this kind and on this scale. The identified parasite peptides have been validated in sera, and a subset of 12 candidates have shown statistically significant increased damage to tick survival (80-100%) [381]. Two of six selected antigens have also shown greater than 70% efficacy in cattle vaccine trials (unpublished).

5.5 Conclusion

This analysis builds on the previous report by Guerrero et al. in 2010 [362], to characterise genomic DNA in the tick *Rhipicelphalus microplus.* The complete secreted extracellular matrix protein gene *papilin* primarily found in basement membranes and essential for embryonic development, was assembled and cDNA sequenced. This is the first reporting in eukaryotes of same strand exon overlap between the sequenced products of *papilin* and *helicase*. Detection of these types of overlaps is a complication for current de-novo gene prediction tools. In a second BAC clone, ribosomal DNA (rDNA) was assembled into three repeat units, the first rRNA assembly in *Rhipicephalinae*, and the first attempt to assemble sequence of the rDNA repeat units and intergenic spacer in arthropods.

In both *papilin* and rRNA, tick specific sites of sequence variation were identified in a detailed comparison of tick *R. microplus* relative to the host *Bos taurus*, in order to identify targets for disrupting the pathogen-host interaction. In addition expression analysis of *papilin* and *helicase* demonstrated striking tissue specific expression in response to sensing the host prior to attachment for feeding.

The two Cot-filtration resources provided a means to estimate the frequency of an element in the context of the whole genome. In order to place the BAC sequences into a whole genome context, the BAC sequences were probed with 454 sequenced Cot 69.56 secs and 696.6 secs DNA. This analysis allowed the representation of specific BAC sequence to be estimated within the respective Cot DNA sequences and thus estimate the frequency of sequence occurrence in the whole *R. microplus* genome.

The BAC, BAC end sequences (BES) and Cot DNA have allowed an in-depth analysis of selected *R. microplus* genomic DNA, and in terms of sequencing towards a whole genome provided a valuable insight into *R. microplus* genomic structure.

In summary the key findings for this chapter are, 1) the bioinformatics approaches used in this chapter have in a complex understudied species identified genomic level detail and genome wide screening of available data 2) BAC analysis identified gene structure, validated for two genes (papilin and serpin), 3) the bioinformatics

approach for high through put identification of vaccine candidates in transcript is the first of this kind and on this scale, 4) Identified candidate peptides have been validated in host sera, and in preliminary Brazilian vaccine trials have shown greater than 70% efficacy in the host against tick, 5) the workflows from this chapter are being adapted to the prediction of peptide candidates involved in wheat allergy/hypersensitivity in humans.

6 Chapter Six - Conclusion

6.1 Thesis contribution to the field of bioinformatics

Bioinformatics approaches underpin the effective analysis of all biological data, influencing data and information at the following levels: 1) capture, storage and collation; 2) access, sharing and integration; 3) Increasing analysis predictive power and generating hypothesis driven analyses. In the framework of this thesis hypothesis, to identify functional genomic targets in diverse disease mechanisms, the developed *in silico* approaches integrated data from diverse resources that had variable data formats, qualities, availability and accessibility constraints, and demonstrated functional predictions for candidate genomic targets in diverse disease cases. The aims of this thesis relative to the case study chapters (light blue) are summarised in Figure 6.1. The major outcomes for the aims and chapters are also shown in Figure 6.1 (dark blue).

In all three case studies (Chapter 3-5) *in silico* approaches were developed that overcame data and analysis constraints, made functional predictions based on comparative analysis and data integration, and identified disease and pathogenic genomic targets. The approaches developed to integrate and mine available data and information, contributed to hypothesis driven disease research in non-model organisms (bacteria and tick), and human.

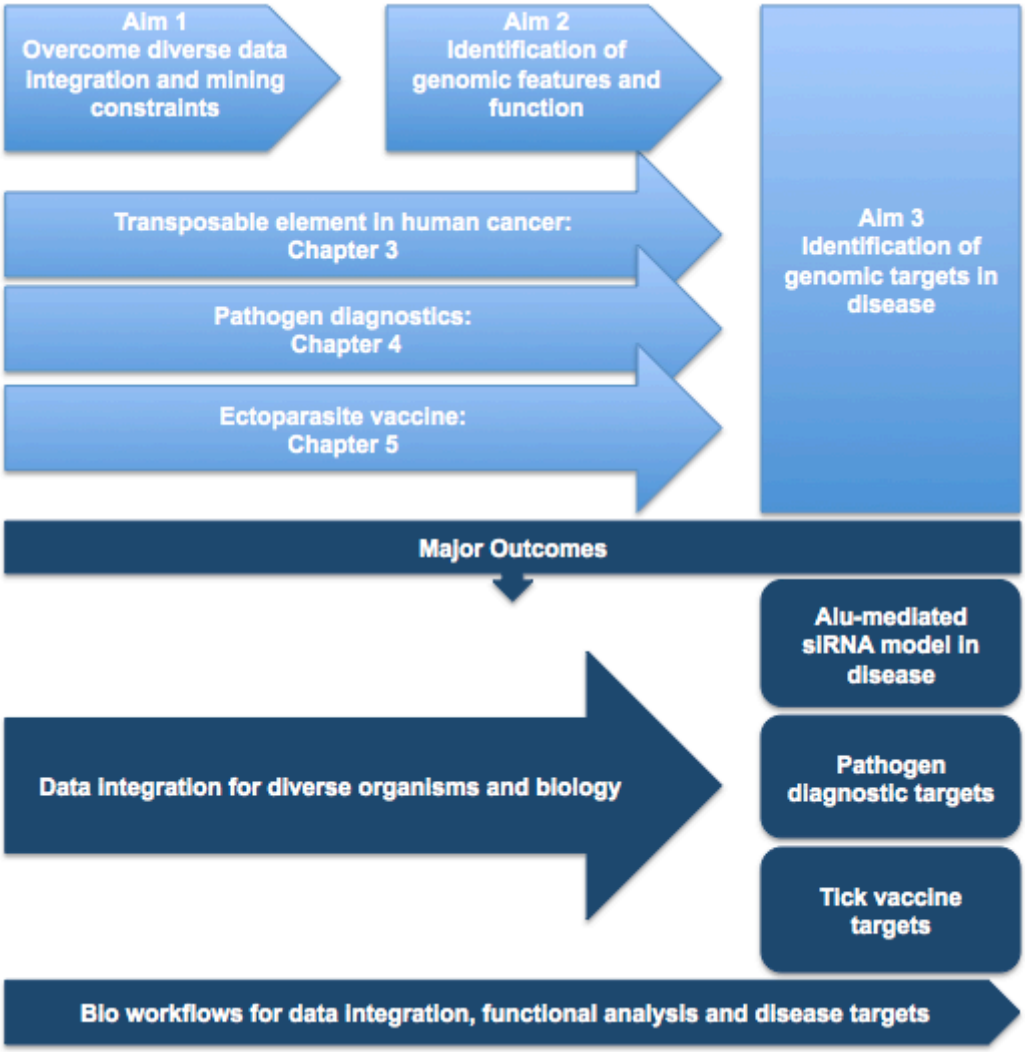


Figure 6.1 Thesis aims (light blue) and related chapter outcomes (dark blue)

Online research environments need to provide to the end user easy access, assembly and reuse of bioinformatics workflows across distributed computing resources. That also simplifies access to computer and data resources and leaves the infrastructure transparent to the end user. In relation to this, the developed and

demonstrated approaches in three diverse biological disease systems contribute to the supply of workflows for hypothesis driven research.

The detailed results from the three case chapters are as follows.

6.2 Case study chapter results

Novel model predictions and findings for the possible role of the human Alu repeat element in human cancer are shown in Table 6.1. The charted organ transcript sequence was assigned to cancer or a normal phenotype and the transcript Alu element content and transcript functions were differentially measured between these two phenotypes. The transcription of Alu was reduced in cancer transcripts and an Alu-mediated siRNA model for the down regulation of Alu containing mRNA proposed. The majority of non-cancerous Alu transcript was of unknown function.

Table 6.1 Case study (Chapter 3), novel predictions and findings for the possible role of Human repeat element in disease. *Homo sapiens* Alu repeat element targets: research outcomes and outputs

Thesis Aims **Develop *in silico* strategies/approaches:**	**Research outputs**	**Research outcomes**	**Extra findings**
To over come various data and analysis constraints. ➢ Integration of transcript and Alu repeat data	➢ Transcript Alu content charted for cancer and normal 'phenotypes' ➢ Transcript Alu content charted for organ source ➢ Alu content measured for human transcript and cancer	➢ Human transcript dataset with cancer and organ source for significant repeat element analyses	➢ Biases tested ➢ Bio-work flow
To make functional predictions based on comparative analyses and data integration ➢ Transcriptome functional analysis	➢ An over representation of non cancerous Alu transcript was found with unknown function ➢ In most organs the transcription of Alu is reduced in cancer.	➢ Functional processes are influenced by Alu content in transcript	➢ Exceptions exist for a few organs
To identify genomic targets ➢ Alu containing transcript functional analysis	➢ Alu-mediated siRNA model for the down regulation of Alu containing mRNA	➢ Alu is a possible candidate target in cancer	➢ Normal tissue with full length Alu are hypothetical in function

Genomic diagnostic analysis findings for *Campylobacter* an important pathogen in cattle venereal disease are shown in Table 6.2. A *Campylobacter* sub species-specific 80Kb genomic sequence that contained plasmid and virulence genes was identified for diagnostics. The candidate diagnostic specificity varied for CFV biovars, as the Type IV Secretary pathway genes are plasmid borne and therefore unstable.

Table 6.2 Case study (Chapter 4), comparative genomic analyses for the identification of gene targets for Campylobacter diagnostics. *Campylobacter fetus* subspecies *venerealis* diagnostics genes for pathogen identification: research outcomes and outputs

Thesis Aims **Develop *in silico* strategies/approaches:**	**Research outputs**	**Research outcomes**	**Extra findings**
To over come various data and analysis constraints in Prokaryote. ➢ Integrate available genomic data to identify genomic elements unique to CFV	➢ CFV pseudo genome assembled. ➢ Comparative genomic analysis of CFV Contigs 1.1 MB (75% complete) to reference genome CFF ➢ CFV 80 Kb unique sequence identified	➢ Unique 80Kb CFV sequence for diagnostics.	➢ CFV assembly reference upheld after complete sequencing of CFV
To make functional predictions based on comparative analyses and data integration ➢ Comparative and functional analysis of genes predictions.	➢ Predicted and functional annotated gene set for CFV. ➢ CFV virulence genes identified. ➢ Comparative analysis of Campylobacter spp.	➢ Type IV Secretary pathway component candidates for PCR analysis	➢ Plasmid and virulence genes found in the 80Kb CFV unique sequence. Including the Type IV secretary pathway genes.
To identify genomic targets ➢ Identify candidates for CFV diagnostics	➢ PCR analysis of 34 candidate genes ➢ 9 candidates specific for CFV subspecies	➢ Diagnostic virulence gene targets for CFV subspecies ➢ New Pfizer CFV biovar genome sequencing project	➢ Candidates' specificity varied for CFV biovars, and compromised diagnostic value. ➢ Type IV Secretary pathway genes are possibly plasmid borne in certain biovars

The findings for the de-novo identification of tick vaccination targets in a large complex ectoparasite un-sequenced genome are detailed in Table 6.3. The available genomic and transcript sequence of *Rhipicephalus microplus* was characterised in depth to support future sequencing efforts and predictions in this organism. BAC sequences were assembled, annotated and genome wide predictions made on feature frequencies. Through the development of a high through put transcript workflow, vaccine candidates were identified that met peptide functional and structural requirements. These peptide vaccine targets have shown greater than 70% efficacy in host trials and are in a patent application.

Table 6.3 Case study (Chapter 5), predicting in *Rhipicephalus microplus* complex genome target genes for parasite control: research outputs and outcomes.

Thesis Aims **Develop *in silico* strategies/approaches:**	**Research outputs**	**Research outcomes**	**Extra findings/outcomes**
To over come various data and analysis constraints ➢ No reference genome ➢ No reference transcriptome ➢ Complex genome (highly repetitive (70%), ➢ No QTL or phenotypic data available	➢ Targeted BAC sequencing based on BES analysis ➢ *De novo* assembly of BAC sequences with BES ➢ *De novo* genomic assembly of 454 Cot DNA (WGS) sequence ➢ BAC Gene prediction ➢ Cot DNA validation of gene predictions	➢ Analysis of incomplete genome BES, BAC and Cot DNA ➢ Two targeted BAC clones sequenced to completion ➢ rRNA and rDNA structure model ➢ Repeat element genome wide frequency estimation method using Cot DNA fractions ➢ Established low genomic sequence homology to Isc (nearest genomic sequence available)	➢ Rmi GSS, WGS GenBank submission ➢ No evidence of Rmi Isc genomic conservation ➢ *De novo* characterization of repeat elements ➢ Ruka element copy number estimated 250K
To make functional predictions based on comparative analyses and data integration ➢ Functional analysis of transcript and expression data	➢ Transcript and BES comparative and functional analysis. ➢ BAC gene prediction and comparative and functional analysis. Two full length gene predictions ➢ Functional annotated transcript set	➢ Complete Rmi BES analysis ➢ Target BES identified in tick feeding ➢ Selection of two target BAC clones for sequencing ➢ Full length papilin cDNA sequenced (Aust. strain) ➢ Transcript set identified for tick feeding and vaccination targets ➢ Isc good reference for Rmi transcriptome	➢ Helicase and papilin exon sharing
To identify genomic targets ➢ Parasite feeding and vaccination targets	➢ Identified targets involved in tick attachment and feeding ➢ Host - pathogen sites of variance in papilin and rRNA protein binding sites	➢ High throughput transcriptome bio-flow for epitope peptide predictions and vaccine design ➢ Peptide vaccine targets have shown >70% efficacy in host trials	➢ Bio-work flows

6.3 Discussion and future work

In this thesis, bioinformatics approaches have been developed and implemented across organisms of diverse origins with variable data and information availability, and disease mechanisms to identify functional genomic candidates. Functional targets have been identified for Alu repeat element in human cancer, host parasite vaccination and the detection by diagnostics for bacterial subspecies known to induce different diseases and symptoms in the host.

In the case of the Human Alu element, novel analysis bioinformatics approaches were implemented for the first time and a model put forward to the research community for validation.

The diagnostic results from the Campylobacter analysis resulted in an ARC linkage project (LP0883837) with industry partners Pfizer Australia and Gribbles Veterinary Pathology to characterise and study Australian isolates (biovars) of *Cfv*. This project will use an integrated genomics sequencing approach, comparative genomics plus molecular and phenotypic screening to improve the understanding of the biology of genital campylobacteriosis in cattle.

The tick analysis from this thesis has directly supported the development of a 'CattleTickBase' resource, toward the sequencing of the *R. microplus* genome. The

findings for the tick investigation (Table 6.3) have provided opportunities for further tick control research and a patent application. As new transcriptomes are sequenced, the time to vaccine candidate identification in pests is reduced by the adoption of the high through put epitope workflow/pipeline developed in this thesis. This pipeline has been selected for wheat to identify immunological targets for allergies in EST libraries. The sequencing of the full-length papilin from BAC sequence contributed a validated gene structure required for training gene prediction tools. Many more sequences of laboratory evidence are needed to establish a tick gene set with known, not *ab initio* predicted, intron-exon structure.

Finally, through the shared bioinformatics approaches used in this thesis, novel functional targets and models in disease can be determined in diverse organisms. These approaches applicable to future online research workflow environments.

6.4 Summary Conclusion

This thesis has developed *in silico* approaches that has delivered:

1. The collation, annotation and integration of available data sets, in three diverse systems each with varying data and integration constraints.
2. The identification of functional genomic targets in diverse biological systems with diverse disease mechanisms.
3. Novel functional analysis in three diverse systems each with varying data and integration constraints.

4. Novel functional genomic targets independent of the disease system in study.
5. High throughput vaccine and diagnostic candidate identification in non-model organisms.

Appendix

Appendix 3.1 Please find excel file ‘appendix3.1.xlsx’ of Alu transcript records on the attached CD.

Appendix 3.2 PostgreSQL database table information

```
dev_goannotation=> \dt
           List of relations
 Schema |     Name      | Type  | Owner
--------+---------------+-------+-------
 public | cdna          | table | paula
 public | gene          | table | paula
 public | gene2go       | table | paula
 public | go            | table | paula
 public | goslim        | table | paula
 public | repeatmasker  | table | paula
 (8 rows)

dev_goannotation=> \d cdna
        Table "public.cdna"
   Column     |  Type   | Modifiers
--------------+---------+-----------
 accession    | text    | not null
 loci         | text    |
 gene_id      | text    |
 tissue       | text    |
 tissue_type  | text    |
 state        | text    |
 cds_start    | integer |
 cds_end      | integer |
 fcfunction   | text    |
 cdna_length  | integer |
Indexes:
    "cdna_pkey" PRIMARY KEY, btree (accession)
Foreign-key constraints:
    "$1" FOREIGN KEY (gene_id) REFERENCES gene(gene_id) ON UPDATE
RESTRICT ON DELETE RESTRICT

dev_goannotation=> \d gene
         Table "public.gene"
      Column       | Type | Modifiers
-------------------+------+-----------
 gene_id           | text | not null
 taxa              | text |
 gene_symbol       | text |
 location          | text |
 gene_description  | text |
Indexes:
```

```
    "gene_pkey" PRIMARY KEY, btree (gene_id)

dev_goannotation=> \d gene2go
     Table "public.gene2go"
    Column    | Type | Modifiers
--------------+------+-----------
 go_id        | text | not null
 gene_id      | text | not null
 go_name      | text |
 go_namespace | text |
 evidence     | text |
Indexes:
    "genego_pkey" PRIMARY KEY, btree (gene_id, go_id)
Foreign-key constraints:
    "$1" FOREIGN KEY (gene_id) REFERENCES gene(gene_id) ON UPDATE
RESTRICT ON DELETE RESTRICT

dev_goannotation=> \d go
      Table "public.go"
  Column   | Type | Modifiers
-----------+------+-----------
 go_id     | text | not null
 name      | text |
 namespace | text |
 go_isa    | text |
Indexes:
    "go_pkey" PRIMARY KEY, btree (go_id)

dev_goannotation=> \d goslim
        Table "public.goslim"
      Column       | Type | Modifiers
-------------------+------+-----------
 go_id             | text | not null
 go_slim_id        | text | not null
 go_slim_name      | text |
 go_slim_namespace | text |
Indexes:
    "go_slim_pk" PRIMARY KEY, btree (go_id, go_slim_id)
Foreign-key constraints:
    "go2slim_fk" FOREIGN KEY (go_id) REFERENCES go(go_id)

dev_goannotation=> \d repeatmasker
        Table "public.repeatmasker"
   Column    |       Type       | Modifiers
-------------+------------------+-----------
 accession   | text             |
 id          | integer          | not null
 score       | double precision |
 percdiv     | double precision |
 percdel     | double precision |
 percin      | double precision |
 querystart  | integer          |
```

```
 queryend    | integer          |
 queryleft   | integer          |
 strand      | text             |
 repeat      | text             |
 class       | text             |
 repeatstart | integer          |
 repeatend   | integer          |
 repeatleft  | integer          |
Indexes:
    "id_pkey" PRIMARY KEY, btree (id)
Foreign-key constraints:
    "repeat2cdna" FOREIGN KEY (accession) REFERENCES cdna(accession)
ON UPDATE RESTRICT ON DELETE RESTRICT
```

Appendix 3.3 Alu-transcript in normal and cancerous tissues statistical analysis crosstabs

Crosstabs

Notes

Output Created		22-NOV-2007 10:48:31
Comments		
Input	**Filter**	<none>
	Weight	Count
	Split File	<none>
	N of Rows in Working Data File	99
Missing Value Handling	**Definition of Missing**	User-defined missing values are treated as missing.
	Cases Used	Statistics for each table are based on all the cases with valid data in the specified range(s) for all variables in each table.
Syntax		CROSSTABS /TABLES=Tissue BY Alu BY Condition /FORMAT= AVALUE TABLES /STATISTIC=CHISQ /CELLS= COUNT /COUNT ROUND CELL .
Resources	**Elapsed Time**	0:00:00.27
	Dimensions Requested	3
	Cells Available	104924

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Tissue * Alu * Condition	106825	100.0%	0	.0%	106825	100.0%

Tissue * Alu * Condition Crosstabulation

Count

Condition			Alu		Total
			no	yes	
bladder	Tissue	disease	380	47	427
		normal	76	34	110
	Total		456	81	537

bone	**Tissue**	**disease**	451	61	512
		normal	255	80	335
	Total		706	141	847
brain	**Tissue**	**disease**	10131	637	10768
		normal	18352	4753	23105
	Total		28483	5390	33873
cervix	**Tissue**	**disease**	2415	171	2586
		normal	201	155	356
	Total		2616	326	2942
colon	**Tissue**	**disease**	2983	469	3452
		normal	893	360	1253
	Total		3876	829	4705
esophagus	**Tissue**	**disease**	158	71	229
		normal	11	9	20
	Total		169	80	249
eye	**Tissue**	**disease**	1203	198	1401
		normal	733	264	997
	Total		1936	462	2398
kidney	**Tissue**	**disease**	203	50	253
		normal	1855	741	2596
	Total		2058	791	2849
liver	**Tissue**	**disease**	135	33	168
		normal	3571	425	3996
	Total		3706	458	4164
lung	**Tissue**	**disease**	2144	222	2366
		normal	1331	580	1911
	Total		3475	802	4277
lymph	**Tissue**	**disease**	932	142	1074
		normal	1652	795	2447
	Total		2584	937	3521
mammary	**Tissue**	**disease**	137	24	161
		normal	309	290	599
	Total		446	314	760
muscle	**Tissue**	**disease**	1061	116	1177
		normal	1000	163	1163
	Total		2061	279	2340

oral	**Tissue**	**disease**	404	241	645
		normal	764	200	964
	Total		1168	441	1609
ovary	**Tissue**	**disease**	204	81	285
		normal	262	39	301
	Total		466	120	586
pancreas	**Tissue**	**disease**	759	77	836
		normal	184	20	204
	Total		943	97	1040
placenta	**Tissue**	**disease**	1485	219	1704
		normal	15468	1284	16752
	Total		16953	1503	18456
prostate	**Tissue**	**disease**	95	28	123
		normal	672	368	1040
	Total		767	396	1163
rectum	**Tissue**	**disease**	102	39	141
		normal	43	49	92
	Total		145	88	233
skin	**Tissue**	**disease**	2705	395	3100
		normal	326	37	363
	Total		3031	432	3463
stomach	**Tissue**	**disease**	36	3	39
		normal	1042	422	1464
	Total		1078	425	1503
testis	**Tissue**	**disease**	741	136	877
		normal	7481	1854	9335
	Total		8222	1990	10212
thyroid	**Tissue**	**disease**	14	0	14
		normal	203	95	298
	Total		217	95	312
uterus	**Tissue**	**disease**	1352	208	1560
		normal	1191	643	1834
	Total		2543	851	3394
embryo	**Tissue**	**disease**	18	2	20
		normal	841	531	1372
	Total		859	533	1392

Chi-Square Tests

Condition		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
bladder	Pearson Chi-Square	27.048(b)	1	.000		
	Continuity Correction(a)	25.517	1	.000		
	Likelihood Ratio	23.457	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	537				
bone	Pearson Chi-Square	20.898(c)	1	.000		
	Continuity Correction(a)	20.045	1	.000		
	Likelihood Ratio	20.441	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	847				
brain	Pearson Chi-Square	1179.042(d)	1	.000		
	Continuity Correction(a)	1177.946	1	.000		
	Likelihood Ratio	1364.484	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	33873				
cervix	Pearson Chi-Square	433.061(e)	1	.000		
	Continuity Correction(a)	429.321	1	.000		
	Likelihood Ratio	301.892	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	2942				
colon	Pearson Chi-Square	145.267(f)	1	.000		
	Continuity Correction(a)	144.225	1	.000		
	Likelihood Ratio	134.622	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	4705				
esophagus	Pearson Chi-Square	1.652(g)	1	.199		
	Continuity Correction(a)	1.073	1	.300		
	Likelihood Ratio	1.572	1	.210		
	Fisher's Exact Test				.217	.150
	N of Valid Cases	249				
eye	Pearson Chi-Square	57.087(h)	1	.000		
	Continuity Correction(a)	56.296	1	.000		
	Likelihood Ratio	56.321	1	.000		

	Fisher's Exact Test				.000	.000
	N of Valid Cases	2398				
kidney	**Pearson Chi-Square**	8.863(i)	1	.003		
	Continuity Correction(a)	8.431	1	.004		
	Likelihood Ratio	9.439	1	.002		
	Fisher's Exact Test				.003	.001
	N of Valid Cases	2849				
liver	**Pearson Chi-Square**	13.362(j)	1	.000		
	Continuity Correction(a)	12.457	1	.000		
	Likelihood Ratio	11.233	1	.001		
	Fisher's Exact Test				.001	.001
	N of Valid Cases	4164				
lung	**Pearson Chi-Square**	305.061(k)	1	.000		
	Continuity Correction(a)	303.686	1	.000		
	Likelihood Ratio	309.072	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	4277				
lymph	**Pearson Chi-Square**	141.876(l)	1	.000		
	Continuity Correction(a)	140.891	1	.000		
	Likelihood Ratio	155.223	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	3521				
mammary	**Pearson Chi-Square**	58.759(m)	1	.000		
	Continuity Correction(a)	57.386	1	.000		
	Likelihood Ratio	65.162	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	760				
muscle	**Pearson Chi-Square**	9.640(n)	1	.002		
	Continuity Correction(a)	9.248	1	.002		
	Likelihood Ratio	9.677	1	.002		
	Fisher's Exact Test				.002	.001
	N of Valid Cases	2340				
oral	**Pearson Chi-Square**	53.634(o)	1	.000		
	Continuity Correction(a)	52.802	1	.000		
	Likelihood Ratio	52.938	1	.000		
	Fisher's Exact Test				.000	.000

	N of Valid Cases	1609				
ovary	**Pearson Chi-Square**	21.498(p)	1	.000		
	Continuity Correction(a)	20.559	1	.000		
	Likelihood Ratio	21.817	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	586				
pancreas	**Pearson Chi-Square**	.068(q)	1	.794		
	Continuity Correction(a)	.016	1	.899		
	Likelihood Ratio	.068	1	.795		
	Fisher's Exact Test				.789	.441
	N of Valid Cases	1040				
placenta	**Pearson Chi-Square**	55.637(r)	1	.000		
	Continuity Correction(a)	54.945	1	.000		
	Likelihood Ratio	48.797	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	18456				
prostate	**Pearson Chi-Square**	7.801(s)	1	.005		
	Continuity Correction(a)	7.250	1	.007		
	Likelihood Ratio	8.286	1	.004		
	Fisher's Exact Test				.005	.003
	N of Valid Cases	1163				
rectum	**Pearson Chi-Square**	15.525(t)	1	.000		
	Continuity Correction(a)	14.455	1	.000		
	Likelihood Ratio	15.474	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	233				
skin	**Pearson Chi-Square**	1.934(u)	1	.164		
	Continuity Correction(a)	1.707	1	.191		
	Likelihood Ratio	2.033	1	.154		
	Fisher's Exact Test				.179	.093
	N of Valid Cases	3463				
stomach	**Pearson Chi-Square**	8.365(v)	1	.004		
	Continuity Correction(a)	7.356	1	.007		
	Likelihood Ratio	10.572	1	.001		
	Fisher's Exact Test				.003	.001
	N of Valid Cases	1503				

testis	**Pearson Chi-Square**	9.684(w)	1	.002		
	Continuity Correction(a)	9.408	1	.002		
	Likelihood Ratio	10.194	1	.001		
	Fisher's Exact Test				.002	.001
	N of Valid Cases	10212				
thyroid	**Pearson Chi-Square**	6.417(x)	1	.011		
	Continuity Correction(a)	5.000	1	.025		
	Likelihood Ratio	10.452	1	.001		
	Fisher's Exact Test				.007	.005
	N of Valid Cases	312				
uterus	**Pearson Chi-Square**	211.809(y)	1	.000		
	Continuity Correction(a)	210.654	1	.000		
	Likelihood Ratio	221.277	1	.000		
	Fisher's Exact Test				.000	.000
	N of Valid Cases	3394				
embryo	**Pearson Chi-Square**	6.873(z)	1	.009		
	Continuity Correction(a)	5.712	1	.017		
	Likelihood Ratio	8.313	1	.004		
	Fisher's Exact Test				.009	.005
	N of Valid Cases	1392				

a Computed only for a 2x2 table

b 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.59.

c 0 cells (.0%) have expected count less than 5. The minimum expected count is 55.77.

d 0 cells (.0%) have expected count less than 5. The minimum expected count is 1713.44.

e 0 cells (.0%) have expected count less than 5. The minimum expected count is 39.45.

f 0 cells (.0%) have expected count less than 5. The minimum expected count is 220.77.

g 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.43.

h 0 cells (.0%) have expected count less than 5. The minimum expected count is 192.08.

i 0 cells (.0%) have expected count less than 5. The minimum expected count is 70.24.

j 0 cells (.0%) have expected count less than 5. The minimum expected count is 18.48.

k 0 cells (.0%) have expected count less than 5. The minimum expected count is 358.34.

l 0 cells (.0%) have expected count less than 5. The minimum expected count is 285.81.

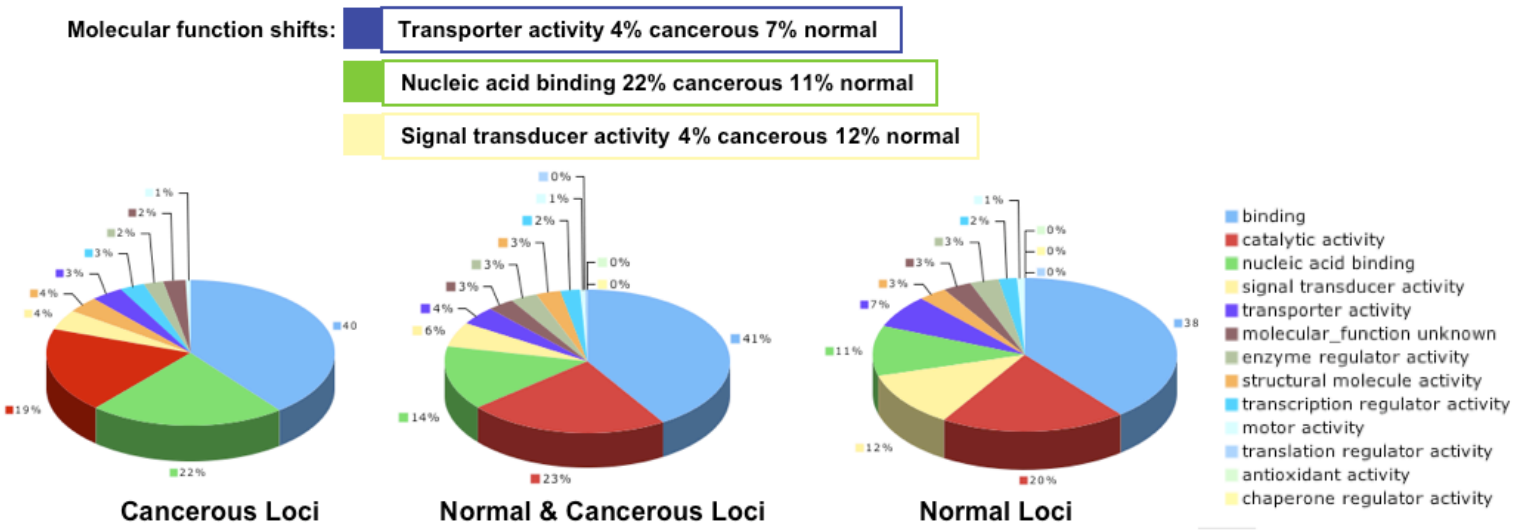
m 0 cells (.0%) have expected count less than 5. The minimum expected count is 66.52.

n 0 cells (.0%) have expected count less than 5. The minimum expected count is 138.67.

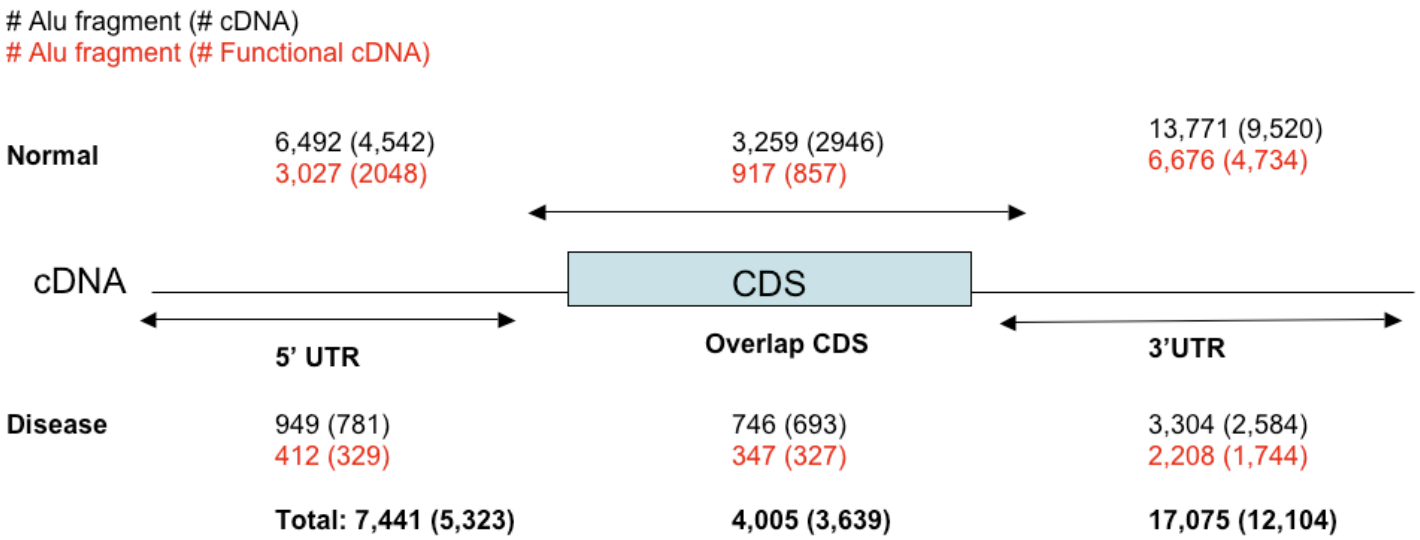
o 0 cells (.0%) have expected count less than 5. The minimum expected count is 176.78.

p 0 cells (.0%) have expected count less than 5. The minimum expected count is 58.36.

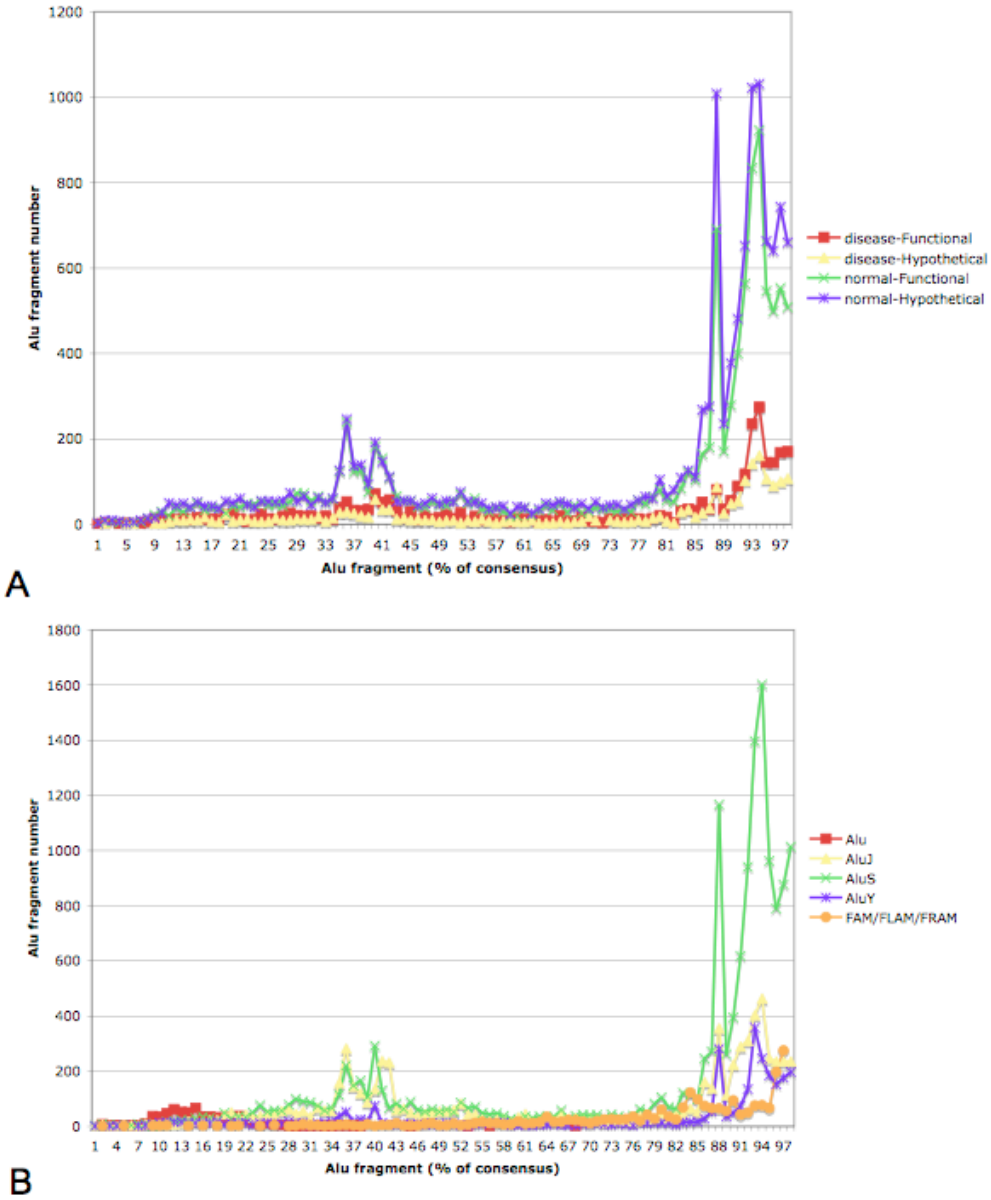
q 0 cells (.0%) have expected count less than 5. The minimum expected count is 19.03.
r 0 cells (.0%) have expected count less than 5. The minimum expected count is 138.77.
s 0 cells (.0%) have expected count less than 5. The minimum expected count is 41.88.
t 0 cells (.0%) have expected count less than 5. The minimum expected count is 34.75.
u 0 cells (.0%) have expected count less than 5. The minimum expected count is 45.28.
v 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.03.
w 0 cells (.0%) have expected count less than 5. The minimum expected count is 170.90.
x 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.26.
y 0 cells (.0%) have expected count less than 5. The minimum expected count is 391.15.
z 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.66.



Appendix 3.4 Gene Ontology for transcript containing Alu elements in Cancerous and Normal tissue.



Appendix Figure 3.5 Alu content in transcript regions for normal and disease tissue



Appendix 3.6 Alu fragment lengths A) in normal and diseased tissue B) Alu families.

Appendix Table 3.7 Subset of full-length Alu-transcript spanning CDS isoforms function exanimation.

Accession	gene_id	loci	cancer	gene_id	gene_symbol	location	gene_description	repeat	isoformProtein
AK022432	GID:950	HIX0018713	normal	GID:950	SCARB2	4q21.1	scavenger receptor class B, member 2	FLAM_A	
AK025390	GID:25926	HIX0014090	normal	GID:25926	NOL11	17q24.2	nucleolar protein 11	FLAM_C	Hypothetical
AK026942	GID:411	HIX0018251	cancer	GID:411	ARSB	5p11-q13	arylsulfatase B	AluSx	Hypothetical
AK024421	GID:56996	HIX0006933	normal	GID:56996	SLC12A9	7q22	solute carrier family 12 (potassium/chloride transporters), member 9	AluSx	
AK024421	GID:56996	HIX0006933	normal	GID:56996	SLC12A9	7q22	solute carrier family 12 (potassium/chloride transporters), member 9	AluJb	
AK024494	GID:56996	HIX0006933	normal	GID:56996	SLC12A9	7q22	solute carrier family 12 (potassium/chloride transporters), member 9	AluSx	
AK024494	GID:56996	HIX0006933	normal	GID:56996	SLC12A9	7q22	solute carrier family 12 (potassium/chloride transporters), member 9	AluSx	
AK054840	GID:152641	HIX0007805	normal	GID:152641	FLJ30277	4q35.1	hypothetical protein FLJ30277	AluSq	Hypothetical
AK055885	GID:1	HIX0015537	normal	GID:1	A1BG	19q13.4	alpha-1-B glycoprotein	AluSq	
AK091237	GID:134266	HIX0005298	cancer	GID:134266	GRPEL2	5q33.1	GrpE-like 2, mitochondrial (E. coli)	FLAM_A	Hypothetical
AK092255	GID:169611	HIX0035002	cancer	GID:169611	OLFML2A	9q33.3	olfactomedin-like 2A	AluSq	Hypothetical
AK094948	GID:286207	HIX0008404	normal	GID:286207	C9orf117	9q34.11	chromosome 9 open reading frame 117	AluSx	Hypothetical
AK000385	GID:8635	HIX0017958	normal	GID:8635	RNASET2	6q27	ribonuclease T2	AluSq	Hypoth
AK098245	GID:5549	HIX0001490	normal	GID:5549	PRELP	1q32	proline/arginine-rich end leucine-rich repeat protein	FLAM_C	Hypoth
BC011823	GID:55652	HIX0010579	disease	GID:55652	FLJ20489	12q13.11	hypothetical protein FLJ20489	AluSx	Hypoth
BC010030	GID:202020	HIX0004117	disease	GID:202020	FLJ39653	4p15.32	hypothetical protein FLJ39653	AluSg	Hypoth
AL162039	GID:92597	HIX0004271	normal	GID:92597	MOBKL1A	4q13.3	MOB1, Mps One Binder kinase activator-like 1A (yeast)	AluSx	Hypoth
AF010144	GID:27308	HIX0028517	normal	GID:27308	AD7C-NTP	1p36	neuronal thread protein	AluSc	Hypoth

							AD7c-NTP		
AK129645	GID:83546	HIX0024449	normal	GID:83546	RTBDN	19p12	retbindin	AluSx	Hypoth
AK124186	GID:196743	HIX0009334	normal	GID:196743	PAOX	10q26.3	polyamine oxidase (exo-N4-amino)	AluSx	
AK125252	GID:2523	HIX0010735	normal	GID:2523	FUT1	19q13.3	fucosyltransferase 1 (galactoside 2-alpha-L-fucosyltransferase, H blood group)	AluSx	Hypoth
AK125657	GID:142679	HIX0002652	disease	GID:142679	DUSP19	2q32.1	dual specificity phosphatase 19	AluSq	Hypoth
AK126660	GID:59271	HIX0027783	normal	GID:59271	C21orf63	21q22.11	chromosome 21 open reading frame 63	AluSq	Hypoth
AK127614	GID:9462	HIX0023758	normal	GID:9462	RASAL2	1q24	RAS protein activator like 2	AluSx	
AF218028	GID:81607	HIX0001228	normal	GID:81607	PVRL4	1q22-q23.2	poliovirus receptor-related 4	FLAM_A	Hypoth
BC009467	GID:158960	HIX0056219	disease	GID:158960	LOC158960	Xq28	hypothetical protein BC009467	FLAM_A	Hypoth
BC024593	GID:11165	HIX0033170	disease	GID:11165	NUDT3	6p21.2	nudix (nucleoside diphosphate linked moiety X)-type motif 3	AluSg	Hypoth
BC037327	GID:22873	HIX0011407	normal	GID:22873	DZIP1	13q32.1	DAZ interacting protein 1	FLAM_C	
BC018643	GID:83716	HIX0020171	disease	GID:83716	CRISPLD2	16q24.1	cysteine-rich secretory protein LCCL domain containing 2	AluSx	Hypoth
BC044913	GID:92597	HIX0004271	normal	GID:92597	MOBKL1A	4q13.3	MOB1, Mps One Binder kinase activator-like 1A (yeast)	AluSx	Hypoth
BC060883	GID:400581	HIX0039163	normal	GID:400581	LOC400581	17p11.2	GRB2-related adaptor protein-like	AluSx	Hypoth
CR627453	GID:57653	HIX0008209	normal	GID:57653	KIAA1529	9q22.33	KIAA1529	AluYb8	Hypoth

Appendix 4.1 *Campylobacter fetus venerealis* sequencing

Whole genome shotgun
Clone size estimate: 1,342,601 bp
Coverage: 4.7 X

Genomic libraries.
Cf1: average insert size of 2 Kb (max 4 Kb)
Cf3: average insert size of 4 Kb (max 6 Kb)
Cf2: average insert size of 6 Kb (max 8 Kb)

Sequencing.
Chemistry: Big-Dye
Number of reads: 13,671
Average read length: 867 bp
Reads by library: 6748 (cf1), 4495 (cf2), 2428 (cf3)
Reads by direction: 7001 forward (51%), 6670 reverse (49%)
Overall base composition: 29.9% A, 18.3% C, 19.5% G, 30.3% T

Assembly.
Contigs: 1187
Contigs > 2 Kb: 273
Singlets: 1335

Appendix Table 4.2 *Cfv* unique open reading frame analysis

ORF	GenBank accession	ContigPosition	hit	description	evalue	pid
campy.fasta.screen.Contig1016.orf00005	ACLG01001016	(2328..2609)	ZP_01998475	radical SAM family protein [Beggiatoa sp. PS] gb|EDN71526.1| radical SAM family protein [Beggiatoa sp. PS]	3.00E-26	54.35
campy.fasta.screen.Contig1018.orf00002	ACLG01001018	rev(282..446)	YP_001405856	hypothetical protein CHAB381_0250 [Campylobacter hominis ATCC BAA-381] gb|ABS51661.1| hypothetical protein CHAB381_0250 [Campylobacter hominis ATCC BAA-381]	3.00E-08	52.17
campy.fasta.screen.Contig1018.orf00003	ACLG01001018	(554..754)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1018.orf00004	ACLG01001018	(1050..1754)	YP_001405854	hypothetical protein CHAB381_0248 [Campylobacter hominis ATCC BAA-381] gb|ABS51459.1| conserved hypothetical protein [Campylobacter hominis ATCC BAA-381]	1.00E-17	42.86
campy.fasta.screen.Contig1021.orf00001	ACLG01001021	(6..647)	ZP_02440650	hypothetical protein CLOSS21_03156 [Clostridium sp. SS2/1] gb|EDS20700.1| hypothetical protein CLOSS21_03156 [Clostridium sp. SS2/1]	9.00E-34	43.53
campy.fasta.screen.Contig1021.orf00002	ACLG01001021	(619..933)	XP_001614429	QF122 antigen, putative [Plasmodium vivax SaI-1] gb|EDL44702.1| QF122 antigen, putative [Plasmodium vivax]	0.16	32.56
campy.fasta.screen.Contig1021.orf00003	ACLG01001021	(936..2081)	YP_393452	HipA-like-like [Sulfurimonas denitrificans DSM 1251] gb|ABB44217.1| HipA-like protein-like protein [Sulfurimonas denitrificans DSM 1251]	1.00E-26	27.88
campy.fasta.screen.Contig1021.orf00004	ACLG01001021	rev(2203..2694)	YP_001405759	Cpp17 [Campylobacter hominis ATCC BAA-381] gb|ABS52419.1| Cpp17 [Campylobacter hominis ATCC BAA-381]	1.00E-14	41.57
campy.fasta.screen.Contig1022.orf00008	ACLG01001022	(3625..3753)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1022.orf00009	ACLG01001022	rev(3778..3915)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1023.orf00001	ACLG01001023	(38..853)	YP_001405783	P-type conjugative transfer protein VirB9 [Campylobacter hominis ATCC BAA-381] gb|ABS51941.1| P-type conjugative transfer protein VirB9 [Campylobacter hominis ATCC BAA-381]	2.00E-79	54.74
campy.fasta.screen.Contig1023.orf00002	ACLG01001023	(840..2075)	YP_001405782	cmgb10 [Campylobacter hominis ATCC BAA-381] gb|ABS51228.1| cmgb10 [Campylobacter hominis ATCC BAA-381]	4.00E-64	38.39
campy.fasta.screen.Contig1023.orf00003	ACLG01001023	(2072..3070)	ACA64441	VirB11 [Campylobacter fetus subsp. venerealis]	1.00E-107	58.64
campy.fasta.screen.Contig1023.orf00004	ACLG01001023	(3074..3334)	EAZ62821	predicted protein [Pichia stipitis CBS 6054]	0.36	40.38
campy.fasta.screen.Contig1023.orf00005	ACLG01001023	(3373..3513)	ACA64444	VirB7/cagT-like protein [Campylobacter fetus subsp. venerealis]	3.00E-08	56.52
campy.fasta.screen.Contig1024.orf00001	ACLG01001024	rev(265..801)	NP_860281	hypothetical protein HH0750 [Helicobacter hepaticus ATCC 51449] gb|AAP77347.1| hypothetical protein HH_0750 [Helicobacter hepaticus ATCC 51449]	2.00E-23	37.43
campy.fasta.screen.Contig1024.orf00002	ACLG01001024	rev(845..1975)	YP_001405857	hypothetical protein CHAB381_0251 [Campylobacter hominis ATCC BAA-381] gb|ABS51011.1| hypothetical protein CHAB381_0251 [Campylobacter hominis ATCC BAA-381]	1.00E-07	29.81
campy.fasta.screen.Contig1024.orf00003	ACLG01001024	rev(1977..2141)	YP_001405856	hypothetical protein CHAB381_0250 [Campylobacter hominis ATCC BAA-381] gb|ABS51661.1| hypothetical protein CHAB381_0250 [Campylobacter hominis ATCC BAA-381]	3.00E-09	60.87
campy.fasta.screen.Contig1024.orf00005	ACLG01001024	rev(2375..2830)	YP_001273095	collagenase, peptidase family U32 [Methanobrevibacter smithii ATCC 35061] gb|ABQ86727.1| collagenase, peptidase family U32 [Methanobrevibacter smithii ATCC 35061]	8.00E-01	26.03
campy.fasta.screen.Contig1030.orf00001	ACLG01001030	rev(1..483)	YP_001405857	hypothetical protein CHAB381_0251 [Campylobacter hominis ATCC BAA-381] gb|ABS51011.1| hypothetical protein CHAB381_0251 [Campylobacter hominis ATCC BAA-381]	7.00E-16	39.60
campy.fasta.screen.Contig1030.orf00002	ACLG01001030	rev(595..753)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1030.orf00003	ACLG01001030	(1249..2022)	YP_001405854	hypothetical protein CHAB381_0248 [Campylobacter hominis ATCC BAA-381] gb|ABS51459.1| conserved hypothetical protein [Campylobacter hominis ATCC BAA-381]	5.00E-36	39.68
campy.fasta.screen.Contig1032.orf00001	ACLG01001032	(73..351)	YP_001467023	type II secretion system protein E [Campylobacter concisus 13826] gb|EAT97402.1| probable pyridine nucleotide-disulfide oxidoreductase YkgC [Campylobacter concisus 13826]	1.00E-11	54.55
campy.fasta.screen.Contig1034.orf00007	ACLG01001034	(2365..2745)	AAO64237	putative putative periplasmic protein Cf0036 [Campylobacter fetus]	3.00E-60	93.65
campy.fasta.screen.Contig1042.orf00001	ACLG01001042	(419..1666)	YP_063409	cpp14 [Campylobacter coli] gb|AAR29498.1| cpp14 [Campylobacter coli]	0	97.70
campy.fasta.screen.Contig1042.orf00002	ACLG01001042	(1653..2465)	YP_247529	hypothetical protein pTet_01 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAX31282.1| pTet01 [Campylobacter jejuni]	1.00E-150	100.00
campy.fasta.screen.Contig1042.orf00003	ACLG01001042	rev(2487..3191)	YP_247530	hypothetical protein pTet_02 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_001004015.1| cpp16 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAX31283.1| pTet02 [Campylobacter jejuni] gb|EAQ71831.1| cpp16 [Campylobacter jejuni subsp. jejuni 81-176]	1.00E-125	94.87
campy.fasta.screen.Contig1042.orf00004	ACLG01001042	rev(3218..4213)	ZP_01072320	Cpp17 [Campylobacter jejuni subsp. jejuni HB93-13] gb|EAQ59637.1| Cpp17 [Campylobacter jejuni subsp. jejuni HB93-13]	1.00E-175	95.68
campy.fasta.screen.Contig1059.orf00004	ACLG01001059	(2068..2316)	ACA64428	bacteriophage P4-like integrase [Campylobacter fetus subsp. venerealis]	1.00E-35	94.87
campy.fasta.screen.Contig1069.orf00001	ACLG01001069	(62..286)	XP_692921	PREDICTED: similar to E3 SUMO-protein ligase PIAS1 (Protein inhibitor of activated STAT protein 1) (Gu-binding protein) (GBP) (RNA helicase II-binding protein) (DEAD/H box-binding protein 1) [Danio rerio]	3.10E-00	40.00
campy.fasta.screen.Contig1069.orf00003	ACLG01001069	rev(283..426)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1069.orf00009	ACLG01001069	rev(2112..3176)	AAO64222	putative ABC transport system permease protein Cf0016 [Campylobacter fetus]	1.00E-179	93.10
campy.fasta.screen.Contig1071.orf00001	ACLG01001071	(54..341)	ACA64449	transposase OrfB [Campylobacter fetus subsp. venerealis]	3.00E-50	100.00
campy.fasta.screen.Contig1071.orf00004	ACLG01001071	(2868..2984)	XP_001682800	hypothetical protein, conserved [Leishmania major] emb|CAJ03671.1| hypothetical protein, conserved [Leishmania major]	6.80E-00	48.39

campy.fasta.screen.Contig1076.orf00001	ACLG01001076	(80..280)	XP_001449172	hypothetical protein GSPATT00016588001 [Paramecium tetraurelia strain d4-2] emb|CAK81775.1| unnamed protein product [Paramecium tetraurelia]	5.20E-00	36.73
campy.fasta.screen.Contig1078.orf00004	ACLG01001078	rev(1900..2013)	ZP_01809557	hypothetical protein Cj8486_0486c [Campylobacter jejuni subsp. jejuni CG8486] gb|EDK21960.1| hypothetical protein Cj8486_0486c [Campylobacter jejuni subsp. jejuni CG8486]	1.00E-04	82.76
campy.fasta.screen.Contig1080.orf00006	ACLG01001080	(3462..3809)	ACA64448	transposase orfA [Campylobacter fetus subsp. venerealis]	6.00E-22	55.34
campy.fasta.screen.Contig1080.orf00007	ACLG01001080	(3830..3946)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1080.orf00008	ACLG01001080	rev(3943..4158)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1082.orf00005	ACLG01001082	(1969..2079)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1082.orf00009	ACLG01001082	(3256..3840)	YP_001176949	phage tail tape measure protein, lambda family [Enterobacter sp. 638] gb|ABP60898.1| phage tail tape measure protein, lambda family [Enterobacter sp. 638]	6.00E-19	34.39
campy.fasta.screen.Contig1082.orf00010	ACLG01001082	rev(4126..4251)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1082.orf00011	ACLG01001082	(4239..4403)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1088.orf00001	ACLG01001088	rev(363..638)	AAZ43839	putative DHH subfamily 1 protein [Mycoplasma synoviae 53]	2.3	32.10
campy.fasta.screen.Contig1088.orf00011	ACLG01001088	rev(6102..6227)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1090.orf00005	ACLG01001090	(1883..2095)	ZP_00368837	formate dehydrogenase, alpha subunit [Campylobacter lari RM2100] gb|EAL55282.1| formate dehydrogenase, alpha subunit [Campylobacter lari RM2100]	2.00E-28	82.86
campy.fasta.screen.Contig1090.orf00008	ACLG01001090	rev(4145..4354)	XP_001142464	PREDICTED: hypothetical protein [Pan troglodytes]	3.9	32.65
campy.fasta.screen.Contig1091.orf00006	ACLG01001091	(2675..3013)	ZP_00367143	chemotaxis regulatory protein Cj1118c [Campylobacter coli RM2228] gb|EAL57047.1| chemotaxis regulatory protein Cj1118c [Campylobacter coli RM2228]	3.00E-51	89.29
campy.fasta.screen.Contig1091.orf00008	ACLG01001091	rev(3776..3994)	XP_001382383	maltose permease [Pichia stipitis CBS 6054] gb|ABN64354.1| maltose permease [Pichia stipitis CBS 6054]	6.7	34.78
campy.fasta.screen.Contig1096.orf00004	ACLG01001096	(2653..2970)	YP_001482271	putative periplasmic protein [Campylobacter jejuni subsp. jejuni 81116] gb|ABV52294.1| putative periplasmic protein [Campylobacter jejuni subsp. jejuni 81116]	2.00E-28	57.84
campy.fasta.screen.Contig1096.orf00005	ACLG01001096	(3127..3315)	YP_178832	hypothetical protein CJE0828 [Campylobacter jejuni RM1221] gb|AAW34613.1| conserved hypothetical protein [Campylobacter jejuni RM1221]	2.00E-16	66.67
campy.fasta.screen.Contig1096.orf00008	ACLG01001096	(4234..4401)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1101.orf00001	ACLG01001101	rev(414..530)	CAO60849	unnamed protein product [Vitis vinifera]	0.62	68.00
campy.fasta.screen.Contig1103.orf00008	ACLG01001103	(4508..5122)	ZP_00369303	oligopeptide ABC transporter, ATP-binding protein, putative [Campylobacter lari RM2100] gb|EAL54469.1| oligopeptide ABC transporter, ATP-binding protein, putative [Campylobacter lari RM2100]	2.00E-21	32.30
campy.fasta.screen.Contig1106.orf00007	ACLG01001106	rev(3575..3823)	YP_581088	protein of unknown function DUF112, transmembrane [Psychrobacter cryohalolentis K5] gb|ABE75604.1| protein of unknown function DUF112, transmembrane [Psychrobacter cryohalolentis K5]	1.00E-14	50.65
campy.fasta.screen.Contig1106.orf00008	ACLG01001106	rev(3920..4150)	XP_543108	PREDICTED: similar to Hypothetical zinc finger protein KIAA1196 [Canis familiaris]	4	25.37
campy.fasta.screen.Contig1106.orf00011	ACLG01001106	(4897..5040)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1113.orf00012	ACLG01001113	rev(4905..5045)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1114.orf00003	ACLG01001114	(1200..1751)	NP_906508	hypothetical protein WS0253 [Wolinella succinogenes DSM 1740] emb|CAE09408.1| hypothetical protein [Wolinella succinogenes]	3.00E-15	33.33
campy.fasta.screen.Contig1114.orf00007	ACLG01001114	rev(3332..4411)	NP_906505	PROBABLE GALACTOSYLTRANSFERASE [Wolinella succinogenes DSM 1740] emb|CAE09405.1| PROBABLE GALACTOSYLTRANSFERASE [Wolinella succinogenes]	7.00E-75	43.63
campy.fasta.screen.Contig1114.orf00008	ACLG01001114	rev(4468..5457)	YP_001467318	glycosyl transferase, group 1 family protein [Campylobacter concisus 13826] gb|EAT99244.1| probable galactosyltransferase [Campylobacter concisus 13826]	1.00E-116	66.67
campy.fasta.screen.Contig1115.orf00005	ACLG01001115	(2654..2815)	ZP_01994533	hypothetical protein DORLON_00518 [Dorea longicatena DSM 13814] gb|EDM63841.1| hypothetical protein DORLON_00518 [Dorea longicatena DSM 13814]	5.2	36.36
campy.fasta.screen.Contig1117.orf00001	ACLG01001117	(79..366)	YP_001355954	hypothetical protein NIS_0483 [Nitratiruptor sp. SB155-2] dbj|BAF69597.1| hypothetical protein [Nitratiruptor sp. SB155-2]	4.00E-18	52.63
campy.fasta.screen.Contig1117.orf00006	ACLG01001117	rev(2049..2399)	CAI79141	C. elegans protein C30G7.3, confirmed by transcript evidence [Caenorhabditis elegans]	0.28	31.00
campy.fasta.screen.Contig1117.orf00007	ACLG01001117	rev(2479..2919)	YP_001922252	mannose-6-phosphate isomerase, class I [Clostridium botulinum E3 str. Alaska E43] gb|ACD52524.1| mannose-6-phosphate isomerase, class I [Clostridium botulinum E3 str. Alaska E43]	0.35	25.00
campy.fasta.screen.Contig1120.orf00003	ACLG01001120	rev(235..1650)	ACA64432	fused VirB3/VirB4 [Campylobacter fetus subsp. venerealis]	0	99.15
campy.fasta.screen.Contig1120.orf00004	ACLG01001120	rev(1827..2381)	ACA64432	fused VirB3/VirB4 [Campylobacter fetus subsp. venerealis]	2.00E-85	91.06
campy.fasta.screen.Contig1120.orf00006	ACLG01001120	(2819..2938)	ZP_00369282	hypothetical protein CLA1237 [Campylobacter lari RM2100] gb|EAL55031.1| hypothetical protein CLA1237 [Campylobacter lari RM2100]	0.62	52.00
campy.fasta.screen.Contig1120.orf00007	ACLG01001120	rev(3042..3212)	ACA64431	hypothetical protein [Campylobacter fetus subsp. venerealis]	7.00E-13	69.64
campy.fasta.screen.Contig1120.orf00009	ACLG01001120	rev(3900..4784)	ACA64428	bacteriophage P4-like integrase [Campylobacter fetus subsp. venerealis]	1.00E-168	99.66
campy.fasta.screen.Contig1124.orf00007	ACLG01001124	(3903..4037)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1126.orf00002	ACLG01001126	rev(1361..1561)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1126.orf00007	ACLG01001126	rev(3339..3443)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1126.orf00008	ACLG01001126	rev(3501..3650)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1126.orf00009	ACLG01001126	(3682..3831)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1139.orf00004	ACLG01001139	(634..3483)	YP_161134	phage-related minor tail protein [Azoarcus sp. EbN1] emb|CAI10233.1| Phage-related minor tail protein	5.00E-38	33.89

				[Azoarcus sp. EbN1]		
campy.fasta.screen.Contig1139.orf00005	ACLG01001139	(3593..3961)	YP_179447	hypothetical protein CJE1461 [Campylobacter jejuni RM1221] gb|AAW35903.1| hypothetical protein CJE1461 [Campylobacter jejuni RM1221]	0.003	35.38
campy.fasta.screen.Contig1140.orf00001	ACLG01001140	(78..248)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1140.orf00002	ACLG01001140	(425..688)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1140.orf00003	ACLG01001140	(685..870)	YP_001405866	transcriptional repressor NrdR [Campylobacter hominis ATCC BAA-381] gb|ABS52054.1| transcriptional repressor NrdR [Campylobacter hominis ATCC BAA-381]	3.00E-10	51.67
campy.fasta.screen.Contig1140.orf00006	ACLG01001140	(2402..2557)	ZP_00372127	hypothetical protein CUP1443 [Campylobacter upsaliensis RM3195] gb|EAL52271.1| hypothetical protein CUP1443 [Campylobacter upsaliensis RM3195]	5.00E-04	57.50
campy.fasta.screen.Contig1140.orf00007	ACLG01001140	(2554..2664)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1141.orf00003	ACLG01001141	(2804..2914)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1141.orf00006	ACLG01001141	(4416..4553)	YP_001467503	inosine-5'-monophosphate dehydrogenase [Campylobacter concisus 13826] gb|EAT98403.1| inosine-5'-monophosphate dehydrogenase [Campylobacter concisus 13826]	3.00E-12	84.09
campy.fasta.screen.Contig1142.orf00014	ACLG01001142	rev(7026..7139)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1143.orf00001	ACLG01001143	rev(20..175)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1149.orf00001	ACLG01001149	rev(85..315)	YP_434682	Transcriptional regulator [Hahella chejuensis KCTC 2396] gb|ABC30257.1| Transcriptional regulator [Hahella chejuensis KCTC 2396]	8.9	44.12
campy.fasta.screen.Contig1149.orf00002	ACLG01001149	rev(512..616)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1156.orf00006	ACLG01001156	rev(2905..3192)	ACA64449	transposase OrfB [Campylobacter fetus subsp. venerealis]	3.00E-50	100.00
campy.fasta.screen.Contig1157.orf00007	ACLG01001157	(3035..3166)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1159.orf00004	ACLG01001159	(3109..5220)	YP_001489487	ferrous iron transport protein B [Arcobacter butzleri RM4018] gb|ABV66818.1| ferrous iron transport protein B [Arcobacter butzleri RM4018]	0	54.69
campy.fasta.screen.Contig1165.orf00001	ACLG01001165	(194..742)	ACA64438	hypothetical protein [Campylobacter fetus subsp. venerealis]	2.00E-86	99.38
campy.fasta.screen.Contig1165.orf00002	ACLG01001165	(743..1684)	ACA64439	VirB9 [Campylobacter fetus subsp. venerealis]	1.00E-155	95.80
campy.fasta.screen.Contig1165.orf00003	ACLG01001165	(1677..2843)	ACA64440	VirB10 [Campylobacter fetus subsp. venerealis]	0	89.18
campy.fasta.screen.Contig1165.orf00004	ACLG01001165	(2840..3490)	ACA64441	VirB11 [Campylobacter fetus subsp. venerealis]	1.00E-111	93.90
campy.fasta.screen.Contig1165.orf00007	ACLG01001165	(3679..3837)	ACA64441	VirB11 [Campylobacter fetus subsp. venerealis]	3.00E-21	96.15
campy.fasta.screen.Contig1165.orf00008	ACLG01001165	(3837..5348)	ACA64442	VirD4 [Campylobacter fetus subsp. venerealis]	0	97.80
campy.fasta.screen.Contig1165.orf00009	ACLG01001165	(5371..5499)	ACA64442	VirD4 [Campylobacter fetus subsp. venerealis]	0.011	57.45
campy.fasta.screen.Contig1165.orf00010	ACLG01001165	(5688..6041)	ACA64442	VirD4 [Campylobacter fetus subsp. venerealis]	1.00E-61	100.00
campy.fasta.screen.Contig1165.orf00011	ACLG01001165	(6045..6197)	ACA64443	hypothetical protein [Campylobacter fetus subsp. venerealis]	1.00E-16	95.45
campy.fasta.screen.Contig1165.orf00012	ACLG01001165	(6241..6753)	ACA64444	VirB7/cagT-like protein [Campylobacter fetus subsp. venerealis]	3.00E-85	100.00
campy.fasta.screen.Contig1165.orf00013	ACLG01001165	(6761..7033)	ACA64445	TrbM-like protein [Campylobacter fetus subsp. venerealis]	2.00E-38	100.00
campy.fasta.screen.Contig1165.orf00014	ACLG01001165	(7197..7415)	ACA64445	TrbM-like protein [Campylobacter fetus subsp. venerealis]	1.00E-38	100.00
campy.fasta.screen.Contig1165.orf00015	ACLG01001165	(7417..8394)	ACA64446	hypothetical protein [Campylobacter fetus subsp. venerealis]	1.00E-159	100.00
campy.fasta.screen.Contig1165.orf00016	ACLG01001165	(8426..8917)	ACA64447	hypothetical protein [Campylobacter fetus subsp. venerealis]	1.00E-46	100.00
campy.fasta.screen.Contig1169.orf00003	ACLG01001169	rev(1734..2288)	YP_664506	hypothetical protein Hac_0704 [Helicobacter acinonychis str. Sheeba] emb|CAJ99507.1| conserved hypothetical protein [Helicobacter acinonychis str. Sheeba]	2.00E-58	58.70
campy.fasta.screen.Contig1169.orf00004	ACLG01001169	rev(2290..2421)	NP_860750	hypothetical protein HH1219 [Helicobacter hepaticus ATCC 51449] gb|AAP77816.1| conserved hypothetical protein [Helicobacter hepaticus ATCC 51449]	1.00E-10	73.81
campy.fasta.screen.Contig1170.orf00001	ACLG01001170	(97..555)	YP_001408866	putative methyltransferase [Campylobacter curvus 525.92] gb|EAU01040.1| putative methyltransferase [Campylobacter curvus 525.92]	5.00E-11	29.22
campy.fasta.screen.Contig1172.orf00014	ACLG01001172	(8548..8694)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1173.orf00016	ACLG01001173	rev(7902..8315)	ZP_01774245	conserved hypothetical protein [Geobacter bemidjiensis Bem] gb|EDJ80432.1| conserved hypothetical protein [Geobacter bemidjiensis Bem]	1.4	33.59
campy.fasta.screen.Contig1181.orf00002	ACLG01001181	rev(611..1357)	YP_001405854	hypothetical protein CHAB381_0248 [Campylobacter hominis ATCC BAA-381] gb|ABS51459.1| conserved hypothetical protein [Campylobacter hominis ATCC BAA-381]	1.00E-39	41.70
campy.fasta.screen.Contig1181.orf00004	ACLG01001181	rev(1645..1845)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1181.orf00005	ACLG01001181	(1953..2117)	YP_001405856	hypothetical protein CHAB381_0250 [Campylobacter hominis ATCC BAA-381] gb|ABS51661.1| hypothetical protein CHAB381_0250 [Campylobacter hominis ATCC BAA-381]	2.00E-08	52.17
campy.fasta.screen.Contig1181.orf00006	ACLG01001181	(2119..4506)	NP_860281	hypothetical protein HH0750 [Helicobacter hepaticus ATCC 51449] gb|AAP77347.1| hypothetical protein HH_0750 [Helicobacter hepaticus ATCC 51449]	1.00E-87	28.85
campy.fasta.screen.Contig1181.orf00007	ACLG01001181	(4717..5736)	NP_860283	hypothetical protein HH0752 [Helicobacter hepaticus ATCC 51449] gb|AAP77349.1| conserved hypothetical protein [Helicobacter hepaticus ATCC 51449]	1.00E-62	44.23
campy.fasta.screen.Contig1181.orf00008	ACLG01001181	(5989..6318)	YP_001405862	hypothetical protein CHAB381_0256 [Campylobacter hominis ATCC BAA-381] gb|ABS51360.1| hypothetical protein CHAB381_0256 [Campylobacter hominis ATCC BAA-381]	3.00E-20	49.53
campy.fasta.screen.Contig1181.orf00010	ACLG01001181	(6396..6809)	YP_011903	hypothetical protein DVU2691 [Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough] gb|AAS97163.1| hypothetical protein DVU_2691 [Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough]	0.004	29.71

campy.fasta.screen.Contig1181.orf00011	ACLG01001181	(6914..7108)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1181.orf00012	ACLG01001181	(7147..7617)	XP_675332	hypothetical protein PB000913.03.0 [Plasmodium berghei strain ANKA] emb|CAI00405.1| conserved hypothetical protein [Plasmodium berghei]	0.21	38.46
campy.fasta.screen.Contig1181.orf00013	ACLG01001181	(7619..8056)	YP_001405865	hypothetical protein CHAB381_0259 [Campylobacter hominis ATCC BAA-381] gb|ABS51884.1| conserved hypothetical protein [Campylobacter hominis ATCC BAA-381]	6.00E-12	48.15
campy.fasta.screen.Contig1181.orf00014	ACLG01001181	(8089..8238)	YP_001405866	transcriptional repressor NrdR [Campylobacter hominis ATCC BAA-381] gb|ABS52054.1| transcriptional repressor NrdR [Campylobacter hominis ATCC BAA-381]	0.001	47.92
campy.fasta.screen.Contig1182.orf00007	ACLG01001182	rev(6932..7066)	AAW56164	Cj81-101 [Campylobacter jejuni]	0.014	80.77
campy.fasta.screen.Contig1182.orf00011	ACLG01001182	(8479..8604)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1182.orf00013	ACLG01001182	(9011..9130)	ZP_01067405	conserved hypothetical protein [Campylobacter jejuni subsp. jejuni CF93-6] ref|ZP_01070403.1| hypothetical protein CJJ26094_0807 [Campylobacter jejuni subsp. jejuni 260.94] ref|ZP_01071660.1| hypothetical protein CJJHB9313_0760 [Campylobacter jejuni subsp. jejuni HB93-13] ref|ZP_01071801.1| hypothetical protein CJJHB9313_0439 [Campylobacter jejuni subsp. jejuni HB93-13] ref|ZP_01072165.1| hypothetical protein CJJHB9313_0037 [Campylobacter jejuni subsp. jejuni HB93-13] gb|EAQ57788.1| conserved hypothetical protein [Campylobacter jejuni subsp. jejuni CF93-6] gb|EAQ58226.1| hypothetical protein CJJ26094_0807 [Campylobacter jejuni subsp. jejuni 260.94] gb|EAQ59714.1| hypothetical protein CJJHB9313_0037 [Campylobacter jejuni subsp. jejuni HB93-13] gb|EAQ59771.1| hypothetical protein CJJHB9313_0439 [Campylobacter jejuni subsp. jejuni HB93-13] gb|EAQ60393.1| hypothetical protein CJJHB9313_0760 [Campylobacter jejuni subsp. jejuni HB93-13]	0.28	94.44
campy.fasta.screen.Contig1182.orf00015	ACLG01001182	(9603..9785)	CAJ30045	conserved hypothetical protein [Magnetospirillum gryphiswaldense]	1.00E-15	68.52
campy.fasta.screen.Contig1182.orf00017	ACLG01001182	(9863..9970)	ABE05728	hypothetical protein UTI89_C0222 [Escherichia coli UTI89] gb|ABE08378.1| hypothetical protein UTI89_C2920 [Escherichia coli UTI89] gb|ABE09159.1| hypothetical protein UTI89_C3719 [Escherichia coli UTI89] gb|ABE09240.1| hypothetical protein UTI89_C3809 [Escherichia coli UTI89] gb|ABE09739.1| hypothetical protein UTI89_C4315 [Escherichia coli UTI89] gb|ABE09854.1| hypothetical protein UTI89_C4441 [Escherichia coli UTI89] gb|ABE09976.1| hypothetical protein UTI89_C4567 [Escherichia coli UTI89]	0.033	61.29
campy.fasta.screen.Contig1182.orf00019	ACLG01001182	(10121..10228)	ZP_02840273	hypothetical protein AchlDRAFT_4885 [Arthrobacter chlorophenolicus A6] gb|EDS59867.1| hypothetical protein AchlDRAFT_4885 [Arthrobacter chlorophenolicus A6]	1.1	70.83
campy.fasta.screen.Contig1182.orf00020	ACLG01001182	(10230..10421)	ZP_01996960	hypothetical protein DORLON_02988 [Dorea longicatena DSM 13814] gb|EDM61652.1| hypothetical protein DORLON_02988 [Dorea longicatena DSM 13814]	1.00E-09	67.44
campy.fasta.screen.Contig1183.orf00005	ACLG01001183	(2893..3348)	ACA64448	transposase orfA [Campylobacter fetus subsp. venerealis]	8.00E-83	100.00
campy.fasta.screen.Contig1183.orf00006	ACLG01001183	(3326..4426)	ACA64449	transposase OrfB [Campylobacter fetus subsp. venerealis]	1.00E-166	85.43
campy.fasta.screen.Contig1184.orf00004	ACLG01001184	rev(2464..2577)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1184.orf00011	ACLG01001184	rev(6614..6763)	YP_001405866	transcriptional repressor NrdR [Campylobacter hominis ATCC BAA-381] gb|ABS52054.1| transcriptional repressor NrdR [Campylobacter hominis ATCC BAA-381]	5.00E-04	47.92
campy.fasta.screen.Contig1184.orf00012	ACLG01001184	rev(6796..7233)	YP_001405865	hypothetical protein CHAB381_0259 [Campylobacter hominis ATCC BAA-381] gb|ABS51884.1| conserved hypothetical protein [Campylobacter hominis ATCC BAA-381]	3.00E-12	48.15
campy.fasta.screen.Contig1184.orf00013	ACLG01001184	rev(7235..7705)	XP_675332	hypothetical protein PB000913.03.0 [Plasmodium berghei strain ANKA] emb|CAI00405.1| conserved hypothetical protein [Plasmodium berghei]	0.21	38.46
campy.fasta.screen.Contig1184.orf00014	ACLG01001184	rev(7744..7938)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1184.orf00016	ACLG01001184	rev(7938..8060)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1184.orf00017	ACLG01001184	rev(8152..8565)	YP_011903	hypothetical protein DVU2691 [Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough] gb|AAS97163.1| hypothetical protein DVU_2691 [Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough]	0.001	29.71
campy.fasta.screen.Contig1184.orf00018	ACLG01001184	rev(8540..8656)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1184.orf00019	ACLG01001184	rev(8649..8783)	YP_001405862	hypothetical protein CHAB381_0256 [Campylobacter hominis ATCC BAA-381] gb|ABS51360.1| hypothetical protein CHAB381_0256 [Campylobacter hominis ATCC BAA-381]	5.1	56.00
campy.fasta.screen.Contig1185.orf00001	ACLG01001185	rev(313..696)	ACA64448	transposase orfA [Campylobacter fetus subsp. venerealis]	4.00E-65	96.85
campy.fasta.screen.Contig1185.orf00004	ACLG01001185	rev(1726..3168)	YP_618234	mobilization protein [Campylobacter jejuni] gb|ABF69295.1| mobilization protein [Campylobacter jejuni]	1.00E-112	53.32
campy.fasta.screen.Contig1185.orf00005	ACLG01001185	rev(3158..3538)	YP_025616	hypothetical protein pAL202_13 [Helicobacter pylori] gb|AAS93847.1| ORF13 [Helicobacter pylori]	3.00E-18	45.83
campy.fasta.screen.Contig1185.orf00006	ACLG01001185	rev(3755..3922)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1185.orf00007	ACLG01001185	(4099..5157)	YP_001405619	putative RepE [Campylobacter hominis ATCC BAA-381] gb|ABS50881.1| putative RepE [Campylobacter hominis ATCC BAA-381]	6.00E-61	40.00
campy.fasta.screen.Contig1185.orf00008	ACLG01001185	rev(5212..5499)	ACA64449	transposase OrfB [Campylobacter fetus subsp. venerealis]	3.00E-50	100.00
campy.fasta.screen.Contig1186.orf00001	ACLG01001186	rev(31..144)	YP_001409071	hypothetical protein CCV52592_0018 [Campylobacter curvus 525.92] gb|EAU00340.1| protein of unknown function [Campylobacter curvus 525.92]	1.00E-05	65.63
campy.fasta.screen.Contig1186.orf00002	ACLG01001186	rev(164..514)	YP_001467495	hypothetical protein CCC13826_0267 [Campylobacter concisus 13826] gb|EAT98338.1| conserved hypothetical protein [Campylobacter concisus 13826]	1.00E-05	30.43
campy.fasta.screen.Contig1186.orf00004	ACLG01001186	rev(594..710)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig1186.orf00005	ACLG01001186	rev(742..1107)	XP_614710	PREDICTED: similar to kinesin-like protein 2 [Bos taurus]	0.008	28.57
campy.fasta.screen.Contig1186.orf00006	ACLG01001186	rev(1104..2174)	YP_861992	two-component system sensor histidine kinase/methyl-esterase/transferase hybrid [Gramella forsetii KT0803] emb|CAL66925.1| two-component system sensor histidine kinase/methyl-esterase/transferase hybrid [Gramella forsetii KT0803]	4.00E-05	27.14

campy.fasta.screen.Contig1186.orf00007	ACLG01001186	rev(2176..2529)	ZP_02178832	hypothetical protein HG1285_12882 [Hydrogenivirga sp. 128-5-R1-1] gb|EDP74401.1| hypothetical protein HG1285_12882 [Hydrogenivirga sp. 128-5-R1-1]	2.00E-12	35.83
campy.fasta.screen.Contig1186.orf00008	ACLG01001186	rev(2522..3160)	YP_001355952	hypothetical protein NIS_0481 [Nitratiruptor sp. SB155-2] dbj|BAF69595.1| hypothetical protein [Nitratiruptor sp. SB155-2]	1.00E-16	40.38
campy.fasta.screen.Contig1186.orf00009	ACLG01001186	rev(3160..4053)	NP_899502	hypothetical protein KVP40.0256 [Vibrio phage KVP40] gb|AAQ64325.1| hypothetical protein KVP40.0256 [Bacteriophage KVP40]	2.00E-06	28.80
campy.fasta.screen.Contig1186.orf00010	ACLG01001186	rev(4067..4513)	XP_999966	PREDICTED: similar to novel member of the keratin associated protein 4 (Krtap4) family [Mus musculus]	0.056	28.43
campy.fasta.screen.Contig1186.orf00011	ACLG01001186	rev(4572..5042)	NP_899500	hypothetical protein KVP40.0254 [Vibrio phage KVP40] gb|AAQ64323.1| hypothetical protein KVP40.0254 [Bacteriophage KVP40]	0.001	29.50
campy.fasta.screen.Contig1186.orf00012	ACLG01001186	rev(5049..5804)	NP_860292	hypothetical protein HH0761 [Helicobacter hepaticus ATCC 51449] gb|AAP77358.1| hypothetical protein HH_0761 [Helicobacter hepaticus ATCC 51449]	7.00E-05	52.08
campy.fasta.screen.Contig1186.orf00013	ACLG01001186	rev(5804..6661)	YP_179446	hypothetical protein CJE1460 [Campylobacter jejuni RM1221] ref|ZP_01099267.1| hypothetical protein CJJ8425_1349 [Campylobacter jejuni subsp. jejuni 84-25] gb|AAW35902.1| hypothetical protein CJE1460 [Campylobacter jejuni RM1221] gb|EAQ94843.1| hypothetical protein CJJ8425_1349 [Campylobacter jejuni subsp. jejuni 84-25]	5.00E-25	27.74
campy.fasta.screen.Contig1186.orf00015	ACLG01001186	rev(6661..9693)	ZP_01099357	hypothetical protein CJJ8425_1350 [Campylobacter jejuni subsp. jejuni 84-25] gb|EAQ94933.1| hypothetical protein CJJ8425_1350 [Campylobacter jejuni subsp. jejuni 84-25]	1.00E-41	24.20
campy.fasta.screen.Contig1187.orf00001	ACLG01001187	rev(351..1796)	YP_447483	member of asn/thr-rich large protein family [Methanosphaera stadtmanae DSM 3091] gb|ABC56840.1| member of asn/thr-rich large protein family [Methanosphaera stadtmanae DSM 3091]	0.002	25.00
campy.fasta.screen.Contig1187.orf00003	ACLG01001187	rev(1797..2384)	NP_860292	hypothetical protein HH0761 [Helicobacter hepaticus ATCC 51449] gb|AAP77358.1| hypothetical protein HH_0761 [Helicobacter hepaticus ATCC 51449]	3.00E-05	55.26
campy.fasta.screen.Contig1187.orf00004	ACLG01001187	rev(2384..3232)	YP_179446	hypothetical protein CJE1460 [Campylobacter jejuni RM1221] ref|ZP_01099267.1| hypothetical protein CJJ8425_1349 [Campylobacter jejuni subsp. jejuni 84-25] gb|AAW35902.1| hypothetical protein CJE1460 [Campylobacter jejuni RM1221] gb|EAQ94843.1| hypothetical protein CJJ8425_1349 [Campylobacter jejuni subsp. jejuni 84-25]	4.00E-30	29.55
campy.fasta.screen.Contig1187.orf00008	ACLG01001187	rev(3235..8508)	ZP_01099357	hypothetical protein CJJ8425_1350 [Campylobacter jejuni subsp. jejuni 84-25] gb|EAQ94933.1| hypothetical protein CJJ8425_1350 [Campylobacter jejuni subsp. jejuni 84-25]	6.00E-43	24.69
campy.fasta.screen.Contig214.orf00005	ACLG01000214	rev(1235..1417)	XP_001301410	hypothetical protein TVAG_481620 [Trichomonas vaginalis G3] gb|EAX88480.1| hypothetical protein TVAG_481620 [Trichomonas vaginalis G3]	8.7	39.02
campy.fasta.screen.Contig419.orf00001	ACLG01000419	(59..217)	YP_001688012	hypothetical protein pK2044_00930 [Klebsiella pneumoniae NTUH-K2044]	2.00E-04	44.68
campy.fasta.screen.Contig419.orf00002	ACLG01000419	rev(294..1226)	ZP_01858414	transcriptional activator [Bacillus sp. SG-1] gb|EDL66246.1| transcriptional activator [Bacillus sp. SG-1]	7.00E-31	29.08
campy.fasta.screen.Contig419.orf00004	ACLG01000419	(1705..1908)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig419.orf00006	ACLG01000419	rev(1889..2116)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig419.orf00007	ACLG01000419	(2145..2387)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig438.orf00001	ACLG01000438	rev(73..327)	ZP_02001316	3-demethylubiquinone-9 3-methyltransferase [Beggiatoa sp. PS] gb|EDN68683.1| 3-demethylubiquinone-9 3-methyltransferase [Beggiatoa sp. PS]	9.00E-05	37.33
campy.fasta.screen.Contig438.orf00002	ACLG01000438	rev(356..490)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig438.orf00006	ACLG01000438	(1293..1418)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig438.orf00007	ACLG01000438	rev(1423..1572)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig438.orf00008	ACLG01000438	(1627..1779)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig683.orf00001	ACLG01000683	rev(135..776)	YP_001407842	copper-translocating P-type ATPase [Campylobacter curvus 525.92] gb|EAU00793.1| copper-translocating P-type ATPase [Campylobacter curvus 525.92]	4.00E-56	56.38
campy.fasta.screen.Contig683.orf00002	ACLG01000683	rev(751..1146)	YP_001407842	copper-translocating P-type ATPase [Campylobacter curvus 525.92] gb|EAU00793.1| copper-translocating P-type ATPase [Campylobacter curvus 525.92]	1.00E-22	46.15
campy.fasta.screen.Contig683.orf00003	ACLG01000683	(1145..1498)	YP_001407842	copper-translocating P-type ATPase [Campylobacter curvus 525.92] gb|EAU00793.1| copper-translocating P-type ATPase [Campylobacter curvus 525.92]	1.00E-13	51.58
campy.fasta.screen.Contig683.orf00004	ACLG01000683	rev(1584..1751)	ZP_01897830	hypothetical protein PE36_09181 [Moritella sp. PE36] gb|EDM67643.1| hypothetical protein PE36_09181 [Moritella sp. PE36]	8.8	34.21
campy.fasta.screen.Contig683.orf00005	ACLG01000683	rev(1779..1928)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig733.orf00001	ACLG01000733	(783..1418)	ZP_00371897	type IV secretion system protein VirB4 [Campylobacter upsaliensis RM3195] gb|EAL52563.1| type IV secretion system protein VirB4 [Campylobacter upsaliensis RM3195]	2.00E-28	38.98
campy.fasta.screen.Contig733.orf00003	ACLG01000733	(1583..1699)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig733.orf00004	ACLG01000733	rev(1859..1969)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig775.orf00007	ACLG01000775	rev(2404..2559)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig791.orf00007	ACLG01000791	(2275..2400)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig797.orf00002	ACLG01000797	rev(439..564)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig797.orf00003	ACLG01000797	(620..751)	ABB00295	mercuric reductase [Alicyclobacillus vulcanalis]	3	45.24
campy.fasta.screen.Contig797.orf00005	ACLG01000797	rev(805..1449)	YP_001408595	probable pyridine nucleotide-disulfide oxidoreductase YkgC [Campylobacter curvus 525.92] gb|EAU01277.1| probable pyridine nucleotide-disulfide oxidoreductase YkgC [Campylobacter curvus 525.92]	6.00E-60	56.56
campy.fasta.screen.Contig806.orf00003	ACLG01000806	rev(1629..1847)	NP_001101458	hypothetical protein LOC313653 [Rattus norvegicus] gb|EDL80882.1| similar to hypothetical protein FLJ32784 (predicted) [Rattus norvegicus]	1.8	33.33

campy.fasta.screen.Contig810.orf00004	ACLG01000810	(1730..1876)	YP_001407659	ribosomal protein S2 [Campylobacter curvus 525.92] gb|EAU00943.2| ribosomal protein S2 [Campylobacter curvus 525.92]	2.4	48.39
campy.fasta.screen.Contig827.orf00010	ACLG01000827	rev(2279..2389)	XP_381451	hypothetical protein FG01275.1 [Gibberella zeae PH-1]	6.9	50.00
campy.fasta.screen.Contig828.orf00003	ACLG01000828	(766..882)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig828.orf00007	ACLG01000828	rev(1822..1974)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig842.orf00006	ACLG01000842	(2281..2478)	XP_001441764	hypothetical protein GSPATT00010566001 [Paramecium tetraurelia strain d4-2] emb|CAK74367.1| unnamed protein product [Paramecium tetraurelia]	4	23.44
campy.fasta.screen.Contig842.orf00007	ACLG01000842	rev(2470..2583)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig846.orf00001	ACLG01000846	rev(200..316)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig846.orf00002	ACLG01000846	rev(343..477)	YP_063413	cpp18 [Campylobacter coli] ref|YP_063462.1| cpp18 [Campylobacter jejuni subsp. jejuni 81-176] ref|ZP_00371109.1| conserved hypothetical protein [Campylobacter coli RM2228] ref|YP_247532.1| hypothetical protein pTet_04 [Campylobacter jejuni subsp. jejuni 81-176] ref|ZP_01072317.1| hypothetical protein CJJHB9313_pTet0004 [Campylobacter jejuni subsp. jejuni HB93-13] ref|YP_001004017.1| cpp18 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAR29502.1| cpp18 [Campylobacter coli] gb|AAR29551.1| cpp18 [Campylobacter jejuni] gb|EAL55772.1| conserved hypothetical protein [Campylobacter coli RM2228] gb|AAX31285.1| pTet04 [Campylobacter jejuni] gb|EAQ59634.1| hypothetical protein CJJHB9313_pTet0004 [Campylobacter jejuni subsp. jejuni HB93-13] gb|EAQ71796.1| cpp18 [Campylobacter jejuni subsp. jejuni 81-176]	1.00E-07	96.43
campy.fasta.screen.Contig846.orf00004	ACLG01000846	(671..979)	YP_063463	cpp19 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_247533.1| hypothetical protein pTet_05 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_001004018.1| cpp19 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAR29552.1| cpp19 [Campylobacter jejuni] gb|AAX31286.1| pTet05 [Campylobacter jejuni] gb|EAQ71786.1| cpp19 [Campylobacter jejuni subsp. jejuni 81-176]	3.00E-34	81.32
campy.fasta.screen.Contig846.orf00005	ACLG01000846	(982..1533)	YP_063464	cpp20 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_247534.1| hypothetical protein pTet_06 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_001004019.1| cpp20 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAR29553.1| cpp20 [Campylobacter jejuni] gb|AAX31287.1| pTet06 [Campylobacter jejuni] gb|EAQ71803.1| cpp20 [Campylobacter jejuni subsp. jejuni 81-176]	4.00E-88	98.20
campy.fasta.screen.Contig846.orf00006	ACLG01000846	(1725..1844)	ZP_00369813	conserved hypothetical protein [Campylobacter lari RM2100] gb|EAL54194.1| conserved hypothetical protein [Campylobacter lari RM2100]	2.00E-11	83.78
campy.fasta.screen.Contig846.orf00008	ACLG01000846	(1933..2655)	YP_063466	cpp22 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_247536.1| hypothetical protein pTet_08 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_001004021.1| cpp22 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAR29555.1| cpp22 [Campylobacter jejuni] gb|AAX31289.1| pTet08 [Campylobacter jejuni] gb|EAQ71815.1| cpp22 [Campylobacter jejuni subsp. jejuni 81-176]	1.00E-109	78.60
campy.fasta.screen.Contig846.orf00010	ACLG01000846	(2806..2952)	ZP_00371123	DNA primase, putative [Campylobacter coli RM2228] gb|EAL55759.1| DNA primase, putative [Campylobacter coli RM2228]	5.00E-14	85.42
campy.fasta.screen.Contig846.orf00011	ACLG01000846	(2965..3120)	YP_063466	cpp22 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_247536.1| hypothetical protein pTet_08 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_001004021.1| cpp22 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAR29555.1| cpp22 [Campylobacter jejuni] gb|AAX31289.1| pTet08 [Campylobacter jejuni] gb|EAQ71815.1| cpp22 [Campylobacter jejuni subsp. jejuni 81-176]	1.00E-13	75.51
campy.fasta.screen.Contig851.orf00002	ACLG01000851	(795..926)	YP_001910368	PARA protein [Helicobacter pylori Shi470] gb|ACD48338.1| PARA protein [Helicobacter pylori Shi470]	0.009	45.24
campy.fasta.screen.Contig851.orf00003	ACLG01000851	(943..1206)	YP_001456649	hypothetical protein CCC13826_0609 [Campylobacter concisus 13826] gb|EAT97555.1| hypothetical protein CCC13826_0609 [Campylobacter concisus 13826]	0.004	32.50
campy.fasta.screen.Contig851.orf00004	ACLG01000851	(1244..1855)	ZP_02949164	sensory transduction histidine kinase [Clostridium butyricum 5521] gb|EDT75845.1| sensory transduction histidine kinase [Clostridium butyricum 5521]	0.028	26.38
campy.fasta.screen.Contig851.orf00005	ACLG01000851	rev(2073..2459)	YP_001456666	HipA domain protein [Campylobacter concisus 13826] gb|EAT97547.1| HipA domain protein [Campylobacter concisus 13826]	3.00E-33	70.83
campy.fasta.screen.Contig851.orf00006	ACLG01000851	rev(2556..2684)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig872.orf00001	ACLG01000872	(94..381)	ACA64449	transposase OrfB [Campylobacter fetus subsp. venerealis]	3.00E-50	100.00
campy.fasta.screen.Contig872.orf00002	ACLG01000872	(411..728)	YP_063476	cpp32 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAR29565.1| cpp32 [Campylobacter jejuni]	2.00E-44	89.80
campy.fasta.screen.Contig872.orf00003	ACLG01000872	(725..1357)	YP_063477	cpp33 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_247546.1| hypothetical protein pTet_18 [Campylobacter jejuni subsp. jejuni 81-176] ref|YP_001004034.1| cpp33 [Campylobacter jejuni subsp. jejuni 81-176] gb|AAR29566.1| cpp33 [Campylobacter jejuni] gb|AAX31299.1| pTet18 [Campylobacter jejuni] gb|EAQ71827.1| cpp33 [Campylobacter jejuni subsp. jejuni 81-176]	1.00E-114	95.02
campy.fasta.screen.Contig872.orf00004	ACLG01000872	(1391..1813)	YP_063428	ssb1 [Campylobacter coli] ref|ZP_00370955.1| single-strand binding protein, putative [Campylobacter coli RM2228] ref|ZP_01072310.1| single-strand binding protein family [Campylobacter jejuni subsp. jejuni HB93-13] gb|AAR29517.1| ssb1 [Campylobacter coli] gb|EAL55921.1| single-strand binding protein, putative [Campylobacter coli RM2228] gb|EAQ59627.1| single-strand binding protein family [Campylobacter jejuni subsp. jejuni HB93-13]	6.00E-75	96.43
campy.fasta.screen.Contig875.orf00001	ACLG01000875	rev(30..170)	ACA64448	transposase orfA [Campylobacter fetus subsp. venerealis]	1.00E-18	97.73
campy.fasta.screen.Contig875.orf00002	ACLG01000875	(274..489)	ACA64432	fused VirB3/VirB4 [Campylobacter fetus subsp. venerealis]	2.00E-32	100.00
campy.fasta.screen.Contig875.orf00003	ACLG01000875	(499..1068)	ACA64433	VirB5 [Campylobacter fetus subsp. venerealis]	3.00E-89	84.66
campy.fasta.screen.Contig875.orf00005	ACLG01000875	(1302..2069)	ACA64434	VirB6 [Campylobacter fetus subsp. venerealis]	1.00E-132	94.12
campy.fasta.screen.Contig899.orf00010	ACLG01000899	rev(3099..3215)	NoHit	NoHit	NoHit	NoHit

campy.fasta.screen.Contig899.orf00012	ACLG01000899	rev(3298..3681)	XP_001886102	predicted protein [Laccaria bicolor S238N-H82] gb|EDR03306.1| predicted protein [Laccaria bicolor S238N-H82]	0.8	51.52
campy.fasta.screen.Contig904.orf00002	ACLG01000904	(1479..1592)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig904.orf00003	ACLG01000904	(1691..1816)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig904.orf00004	ACLG01000904	rev(1824..1961)	YP_001490609	phage repressor protein, putative [Arcobacter butzleri RM4018] gb|ABV67939.1| phage repressor protein, putative [Arcobacter butzleri RM4018]	0.072	40.48
campy.fasta.screen.Contig907.orf00006	ACLG01000907	(2069..2176)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig909.orf00004	ACLG01000909	(1457..1636)	YP_627641	carbamoyl phosphate synthase large subunit [Helicobacter pylori HPAG1] gb|ABF84967.1| carbamoyl-phosphate synthase [Helicobacter pylori HPAG1]	0.071	40.00
campy.fasta.screen.Contig909.orf00005	ACLG01000909	(1804..1932)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig914.orf00001	ACLG01000914	rev(851..985)	YP_001696973	Stage V sporulation protein AD [Lysinibacillus sphaericus C3-41] gb|ACA38843.1| Stage V sporulation protein AD [Lysinibacillus sphaericus C3-41]	1.00E-07	84.85
campy.fasta.screen.Contig914.orf00003	ACLG01000914	(1055..1324)	YP_146708	stage V sporulation protein AE [Geobacillus kaustophilus HTA426] dbj|BAD75140.1| stage V sporulation protein AE [Geobacillus kaustophilus HTA426]	2.00E-37	83.91
campy.fasta.screen.Contig914.orf00004	ACLG01000914	rev(2271..2468)	YP_001696942	alpha-amylase [Lysinibacillus sphaericus C3-41] gb|ACA38812.1| alpha-amylase [Lysinibacillus sphaericus C3-41]	2.00E-08	40.32
campy.fasta.screen.Contig917.orf00001	ACLG01000917	(286..1302)	YP_001482814	hypothetical protein C8J_1238 [Campylobacter jejuni subsp. jejuni 81116] gb|ABV52837.1| hypothetical protein C8J_1238 [Campylobacter jejuni subsp. jejuni 81116]	1.00E-102	54.63
campy.fasta.screen.Contig930.orf00006	ACLG01000930	rev(2933..3175)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig953.orf00009	ACLG01000953	rev(2872..3000)	ACA64449	transposase OrfB [Campylobacter fetus subsp. venerealis]	2.00E-08	83.33
campy.fasta.screen.Contig953.orf00010	ACLG01000953	(2981..3124)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig955.orf00005	ACLG01000955	(3239..3460)	Q705F4	Replication protein E1 (ATP-dependent helicase E1) emb|CAF05687.1| E1 protein [Bovine papillomavirus - 6]	6.9	32.76
campy.fasta.screen.Contig958.orf00001	ACLG01000958	(26..532)	YP_001405788	cmgb3/4 [Campylobacter hominis ATCC BAA-381] gb|ABS51869.1| cmgb3/4 [Campylobacter hominis ATCC BAA-381]	1.00E-33	50.63
campy.fasta.screen.Contig958.orf00002	ACLG01000958	(534..1277)	ACA64433	VirB5 [Campylobacter fetus subsp. venerealis]	7.00E-15	23.51
campy.fasta.screen.Contig958.orf00003	ACLG01000958	(1279..2055)	ZP_00371900	TrbL/VirB6 plasmid conjugal transfer protein [Campylobacter upsaliensis RM3195] gb|EAL52566.1| TrbL/VirB6 plasmid conjugal transfer protein [Campylobacter upsaliensis RM3195]	3.00E-08	25.42
campy.fasta.screen.Contig958.orf00004	ACLG01000958	(2036..2233)	XP_001883804	predicted protein [Laccaria bicolor S238N-H82] gb|EDR05700.1| predicted protein [Laccaria bicolor S238N-H82]	1.4	36.59
campy.fasta.screen.Contig960.orf00003	ACLG01000960	(1429..1671)	ACA64448	transposase orfA [Campylobacter fetus subsp. venerealis]	1.00E-37	97.50
campy.fasta.screen.Contig960.orf00005	ACLG01000960	(2339..2506)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig963.orf00011	ACLG01000963	rev(3977..4141)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig965.orf00003	ACLG01000965	(1513..1617)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig965.orf00007	ACLG01000965	(3886..4008)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig965.orf00008	ACLG01000965	(4118..4243)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig965.orf00009	ACLG01000965	(4255..4434)	YP_372034	hypothetical protein Bcep18194_B1276 [Burkholderia sp. 383] gb|ABB11390.1| hypothetical protein Bcep18194_B1276 [Burkholderia sp. 383]	3.9	41.67
campy.fasta.screen.Contig974.orf00003	ACLG01000974	rev(1173..1280)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig974.orf00006	ACLG01000974	rev(2848..2982)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig975.orf00001	ACLG01000975	rev(83..313)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig977.orf00005	ACLG01000977	rev(2365..2481)	NoHit	NoHit	NoHit	NoHit
campy.fasta.screen.Contig983.orf00001	ACLG01000983	(573..860)	ACA64449	transposase OrfB [Campylobacter fetus subsp. venerealis]	3.00E-50	100.00
campy.fasta.screen.Contig991.orf00001	ACLG01000991	(234..488)	ACA64455	hypothetical protein [Campylobacter fetus subsp. venerealis]	6.00E-22	100.00
campy.fasta.screen.Contig991.orf00002	ACLG01000991	(485..700)	ACA64456	helicase domain protein [Campylobacter fetus subsp. venerealis]	2.00E-33	100.00
campy.fasta.screen.Contig991.orf00003	ACLG01000991	(903..1046)	ACA64457	putative transcriptional regulator [Campylobacter fetus subsp. venerealis]	1.00E-16	100.00
campy.fasta.screen.Contig991.orf00004	ACLG01000991	(1178..1513)	ACA64457	putative transcriptional regulator [Campylobacter fetus subsp. venerealis]	5.00E-59	98.20
campy.fasta.screen.Contig991.orf00005	ACLG01000991	(1581..1898)	ACA64458	hypothetical protein [Campylobacter fetus subsp. venerealis]	2.00E-55	100.00
campy.fasta.screen.Contig991.orf00006	ACLG01000991	(1923..2333)	ACA64459	ParA [Campylobacter fetus subsp. venerealis]	7.00E-61	97.62
campy.fasta.screen.Contig993.orf00001	ACLG01000993	(14..157)	CAL15010	transposase A [Campylobacter fetus subsp. venerealis] emb|CAL15014.1| transposase A [Campylobacter fetus subsp. venerealis] emb|CAJ98526.1| transposase A [Campylobacter fetus subsp. venerealis]	1.00E-07	96.67
campy.fasta.screen.Contig993.orf00002	ACLG01000993	(144..1436)	CAL15011	transposase B [Campylobacter fetus subsp. venerealis] emb|CAL15015.1| transposase B [Campylobacter fetus subsp. venerealis] emb|CAJ98527.1| tnpB; transposase B [Campylobacter fetus subsp. venerealis]	0	94.42
campy.fasta.screen.Contig999.orf00003	ACLG01000999	(2181..2414)	XP_001561687	hypothetical protein LbrM03_V2.0970 [Leishmania braziliensis MHOM/BR/75/M2904] emb|CAM36833.1| hypothetical protein [Leishmania braziliensis]	8.9	44.19
campy.fasta.screen.Contig993.orf00003	ACLG01000993	rev(1509..1697)	YP_892772	anaerobic C4-dicarboxylate transporter [Campylobacter fetus subsp. fetus 82-40] gb|ABK82413.1| anaerobic C4-dicarboxylate transporter DcuA [Campylobacter fetus subsp. fetus 82-40]	4.00E-21	92.98245614
campy.fasta.screen.Contig993.orf00004	ACLG01000993	(1705..2493)	YP_892772	anaerobic C4-dicarboxylate transporter [Campylobacter fetus subsp. fetus 82-40] gb|ABK82413.1| anaerobic C4-dicarboxylate transporter DcuA [Campylobacter fetus subsp. fetus 82-40]	1.00E-130	93.51145038

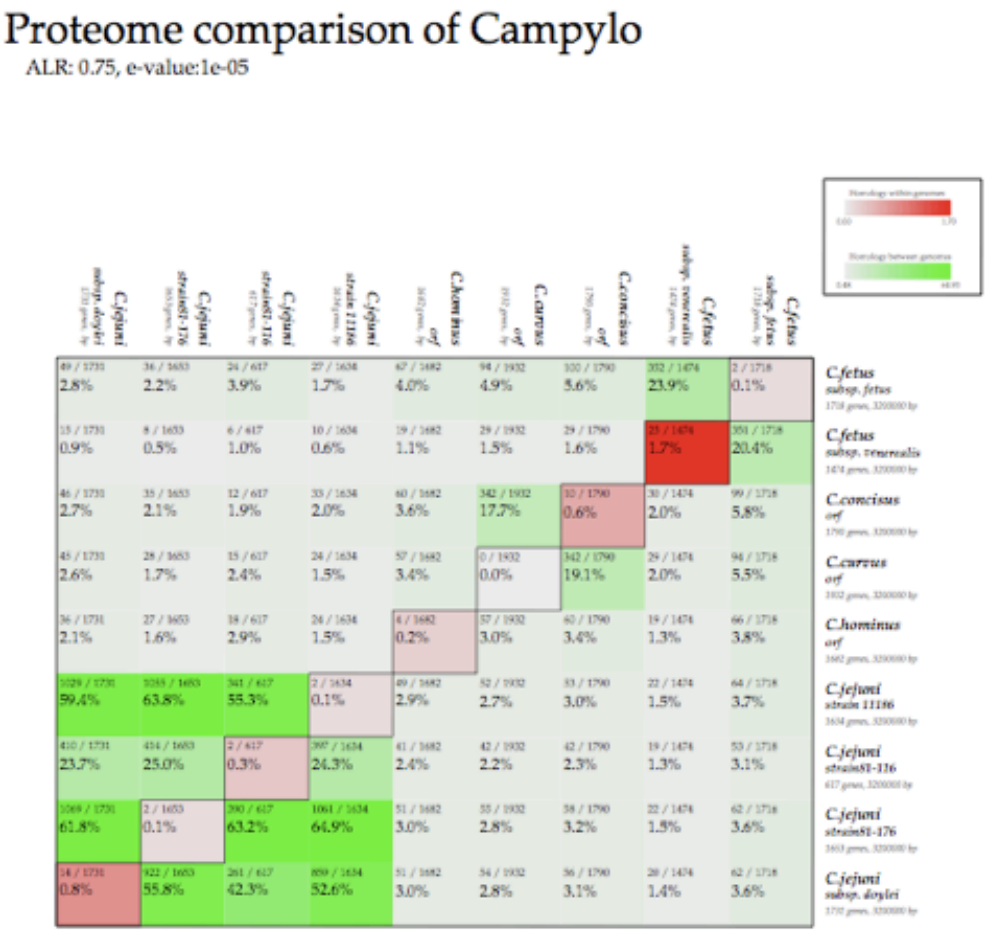
campy.fasta.screen.Contig993.orf00006	ACLG01000993	rev(2795..2968)	YP_892773	methyl-accepting chemotaxis protein [Campylobacter fetus subsp. fetus 82-40] gb|ABK81801.1| methyl-accepting chemotaxis protein [Campylobacter fetus subsp. fetus 82-40]	2.00E-04	64.2857 1429

Appendix Table 4.3 CFV virulence open reading frames selected for PCR

ORF	GenBank accession	ContigPosition	Protein	description	evalue	Pid
campy.fasta.screen.Contig1006.orf00004	ACLG01001006	(2369..2764)	YP_892844	membrane protein [Campylobacter fetus subsp. fetus 82-40] sp|A0RRL1|Y1725_CAMFF UPF0059 membrane protein CFF8240_1725 gb|ABK83235.1| membrane protein [Campylobacter fetus subsp. fetus 82-40]	3.00E-63	100
campy.fasta.screen.Contig1013.orf00003	ACLG01001013	rev(516..1823)	YP_892642	flagellar biosynthesis regulator FlhF [Campylobacter fetus subsp. fetus 82-40] gb|ABK82686.1| flagellar biosynthesis protein [Campylobacter fetus subsp. fetus 82-40]	0	100
campy.fasta.screen.Contig1023.orf00001	ACLG01001023	(38..853)	YP_001405783	P-type conjugative transfer protein VirB9 [Campylobacter hominis ATCC BAA-381] gb|ABS51941.1| P-type conjugative transfer protein VirB9 [Campylobacter hominis ATCC BAA-381]	2.00E-79	54.74452555
campy.fasta.screen.Contig1023.orf00002	ACLG01001023	(840..2075)	YP_001405782	cmgb10 [Campylobacter hominis ATCC BAA-381] gb|ABS51228.1| cmgb10 [Campylobacter hominis ATCC BAA-381]	4.00E-64	38.38862559
campy.fasta.screen.Contig1023.orf00003	ACLG01001023	(2072..3070)	ACA64441	VirB11 [Campylobacter fetus subsp. venerealis]	1.00E-107	58.64197531
campy.fasta.screen.Contig1034.orf00010	ACLG01001034	rev(3516..4052)	YP_891650	putative putative two-component sensor [Campylobacter fetus subsp. fetus 82-40] gb|ABK81714.1| putative putative two-component sensor [Campylobacter fetus subsp. fetus 82-40]	2.00E-84	92.13483146
campy.fasta.screen.Contig1034.orf00012	ACLG01001034	rev(4039..4239)	YP_891651	response regulator ompr [Campylobacter fetus subsp. fetus 82-40] gb|AAO64235.1| putative two-component regulator Cf0034 [Campylobacter fetus] gb|ABK82888.1| response regulator ompr [Campylobacter fetus subsp. fetus 82-40]	4.00E-27	98.33333333
campy.fasta.screen.Contig1037.orf00001	ACLG01001037	(118..993)	YP_891920	haemolysin secretion/activation protein ShlB/FhaC/HecB [Campylobacter fetus subsp. fetus 82-40] gb|ABK82461.1| haemolysin secretion/activation protein ShlB/FhaC/HecB [Campylobacter fetus subsp. fetus 82-40]	1.00E-164	98.96907216
campy.fasta.screen.Contig1040.orf00001	ACLG01001040	rev(186..842)	YP_891447	histidine kinase [Campylobacter fetus subsp. fetus 82-40] gb|ABK81743.1| histidine kinase [Campylobacter fetus subsp. fetus 82-40]	5.00E-72	78.91891892
campy.fasta.screen.Contig1047.orf00002	ACLG01001047	(379..1509)	YP_892016	sensor histidine kinase [Campylobacter fetus subsp. fetus 82-40] gb|ABK82841.1| sensor histidine kinase [Campylobacter fetus subsp. fetus 82-40]	0	96.52406417
campy.fasta.screen.Contig1083.orf00002	ACLG01001083	(322..882)	YP_891727	sensory trasnduction histidine kinase [Campylobacter fetus subsp. fetus 82-40] gb|ABK81914.1| sensory trasnduction histidine kinase [Campylobacter fetus subsp. fetus 82-40]	3.00E-99	99.46236559
campy.fasta.screen.Contig1095.orf00004	ACLG01001095	rev(1376..1741)	YP_891326	iron ABC transporter, permease protein [Campylobacter fetus subsp. fetus 82-40] gb|ABK81910.1| iron ABC transporter, permease protein [Campylobacter fetus subsp. fetus 82-40]	2.00E-60	100
campy.fasta.screen.Contig1120.orf00004	ACLG01001120	rev(1827..2381)	ACA64432	fused VirB3/VirB4 [Campylobacter fetus subsp. venerealis]	2.00E-85	91.06145251
campy.fasta.screen.Contig1143.orf00003	ACLG01001143	(2025..3161)	YP_891345	hypothetical protein CFF8240_0135 [Campylobacter fetus subsp. fetus 82-40] gb|ABK81757.1| conserved hypothetical protein [Campylobacter fetus subsp. fetus 82-40]	0	100
campy.fasta.screen.Contig1154.orf00003	ACLG01001154	rev(1943..3277)	YP_891313	mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase [Campylobacter fetus subsp. fetus 82-40] gb|ABK83083.1| mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase [Campylobacter fetus subsp. fetus 82-40]	0	97.2972973
campy.fasta.screen.Contig1155.orf00004	ACLG01001155	rev(3297..4856)	YP_892764	flagellin [Campylobacter fetus subsp. fetus 82-40] gb|ABK83217.1| flagellin B [Campylobacter fetus subsp. fetus 82-40]	0	90.01919386
campy.fasta.screen.Contig1165.orf00002	ACLG01001165	(743..1684)	ACA64439	VirB9 [Campylobacter fetus subsp. venerealis]	1.00E-155	95.8041958
campy.fasta.screen.Contig1165.orf00004	ACLG01001165	(2840..3490)	ACA64441	VirB11 [Campylobacter fetus subsp. venerealis]	1.00E-111	93.89671362
campy.fasta.screen.Contig1165.orf00008	ACLG01001165	(3837..5348)	ACA64442	VirD4 [Campylobacter fetus subsp. venerealis]	0	97.80439122
campy.fasta.screen.Contig1172.orf00010	ACLG01001172	rev(5136..5933)	YP_891469	flagellar assembly protein H [Campylobacter fetus subsp. fetus 82-40] gb|ABK82698.1| flagellar assembly protein [Campylobacter fetus subsp. fetus 82-40]	1.00E-132	94.33962264
campy.fasta.screen.Contig733.orf00001	ACLG01000733	(783..1418)	ZP_00371897	type IV secretion system protein VirB4 [Campylobacter upsaliensis RM3195] gb|EAL52563.1| type IV secretion system protein VirB4 [Campylobacter upsaliensis RM3195]	2.00E-28	38.98305085
campy.fasta.screen.Contig875.orf00005	ACLG01000875	(1302..2069)	ACA64434	VirB6 [Campylobacter fetus subsp. venerealis]	1.00E-132	94.11764706
campy.fasta.screen.Contig878.orf00002	ACLG01000878	(1087..1713)	YP_891989	flagellar basal body L-ring protein [Campylobacter fetus subsp. fetus 82-40] gb|ABK81968.1| flagellar L-ring protein FlgH [Campylobacter fetus subsp. fetus 82-40]	4.00E-87	81.25
campy.fasta.screen.Contig927.orf00001	ACLG01000927	rev(840..1367)	YP_892122	cytolethal distending toxin A [Campylobacter fetus subsp. fetus 82-40] gb|ABK82550.1| cytolethal distending toxin A [Campylobacter fetus subsp. fetus 82-40]	1.00E-100	98.84393064
campy.fasta.screen.Contig927.orf00002	ACLG01000927	rev(1838..2281)	YP_892123	cytolethal distending toxin A/C family [Campylobacter fetus subsp. fetus 82-40] gb|ABK82032.1| cytolethal distending toxin A/C family [Campylobacter fetus subsp. fetus 82-40]	3.00E-77	100
campy.fasta.screen.Contig927.orf00003	ACLG01000927	rev(2309..2725)	YP_892124	cytolethal distending toxin [Campylobacter fetus subsp. fetus 82-40] gb|ABK82936.1| cytolethal distending toxin [Campylobacter fetus subsp. fetus 82-40]	7.00E-64	91.30434783
campy.fasta.screen.Contig988.orf00001	ACLG01000988	(77..295)	YP_892653	Omp18 [Campylobacter fetus subsp. fetus 82-40] gb|ABK83275.1| Omp18 [Campylobacter fetus subsp. fetus 82-40]	3.00E-35	100
campy.fasta.screen.Contig992.orf00007	ACLG01000992	rev(3393..3611)	YP_892377	fibronectin type III domain protein [Campylobacter fetus subsp. fetus 82-40] gb|ABK81982.1| fibronectin type III domain protein [Campylobacter fetus subsp. fetus 82-40]	1.00E-32	97.22222222
campy.fasta.screen.Contig995.orf00005	ACLG01000995	rev(2271..2648)	YP_891632	putative two-component regulator [Campylobacter fetus subsp. fetus 82-40] gb|ABK83078.1| putative two-component regulator [Campylobacter fetus subsp. fetus 82-40]	7.00E-56	98.16513761

campy.fasta.screen.Contig995.orf00006	ACLG01000995	rev(2645..2785)	YP_891631	His Kinase A (phosphoacceptor) domain protein [Campylobacter fetus subsp. fetus 82-40] gb|ABK82045.1| His Kinase A (phosphoacceptor) domain protein [Campylobacter fetus subsp. fetus 82-40]	9.00E-18	100

Appendix 4.4 Please find the excel file ‘appendix4.4.xlsx’, worksheet ‘cfv_summary’ on the attached CD. This worksheet contains 1147 *Campylobacter fetus* subspecies v*enerialis* open reading frame alignment results to datasets COG (purple), KEGG (blue), Protein (aqua), conserved domain (green) and STRING (yellow).



Appendix 4.5 *Campylobacter* spp. protein matrix analysis

Appendix Table 5.1 BAC predicted gene annotation

Rmi_gene	classified	locus	description	evalue	pid	frame	COG	KEGG	score
BM-004-A11.gene1	BM-004-A11.gene1		Competence domain containing protein	1.00E-04	19% (32/164)	1			95
BM-004-A11.gene10	BM-004-A11.gene10		PREDICTED: similar to LReO_3 n=1 Tax=Strongylocentrotus purpuratus RepID=UPI0000E4971F	3.00E-12	33% (43/127)	2			185
BM-004-A11.gene11	BM-004-A11.gene11		PREDICTED: similar to GA18855-PA n=1 Tax=Tribolium castaneum RepID=UPI0000D56A7F	1.00E-49	39% (110/281)	1			515
BM-004-A11.gene12			unclassified						
BM-004-A11.gene13	BM-004-A11.gene13		Reverse transcriptase n=2 Tax=Ostrinia nubilalis RepID=A4KWG0_OSTNU	4.00E-27	27% (98/356)	1	Replication, recombination and repair		316
BM-004-A11.gene14	BM-004-A11.gene14		PREDICTED: similar to pol-like protein n=1 Tax=Nasonia vitripennis RepID=UPI00015B43EC	2.00E-15	23% (109/460)	1			215
BM-004-A11.gene15	BM-004-A11.gene15		PREDICTED: similar to endonuclease/reverse transcriptase n=1 Tax=Strongylocentrotus purpuratus RepID=UPI0000587485	7.00E-25	35% (62/175)	4			302
BM-004-A11.gene2	BM-004-A11.gene2		Putative gag protein n=1 Tax=Philodina roseola RepID=A1YGR8_9BILA	1.00E-11	29% (37/127)	1			181
BM-004-A11.gene3			unclassified						
BM-004-A11.gene4			unclassified						
BM-004-A11.gene5			unclassified						
BM-004-A11.gene6	BM-004-A11.gene6		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	1.00E-101	34% (207/606)	1			962
BM-004-A11.gene7	BM-004-A11.gene7		PREDICTED: similar to pol polyprotein n=2 Tax=Danio rerio RepID=UPI000175F6E6	4.00E-42	39% (87/219)	1			444
BM-004-A11.gene8	BM-004-A11.gene8		Polyprotein of retroviral origin, putative (Fragment) n=1 Tax=Ixodes scapularis RepID=B7QAL0_IXOSC	2.00E-41	67% (80/119)	2			440
BM-004-A11.gene9			unclassified						
BM-005-B21.gene1			unclassified						
BM-005-B21.gene10			unclassified						
BM-005-B21.gene11	BM-005-B21.gene11		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	1.00E-44	30% (144/471)	1			470
BM-005-B21.gene12			unclassified						
BM-005-B21.gene2			unclassified						
BM-005-B21.gene3			unclassified						
BM-005-B21.gene4	BM-005-B21.gene4		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A3B	4.00E-95	33% (201/603)	1			906
BM-005-B21.gene5	BM-005-B21.gene5		PREDICTED: similar to putative reverse transcriptase n=1 Tax=Strongylocentrotus purpuratus RepID=UPI0000E49809	1.00E-12	39% (41/103)	1			186
BM-005-B21.gene6	BM-005-B21.gene6		PREDICTED: receptor for egg jelly 4-like n=1 Tax=Saccoglossus kowalevskii RepID=UPI0001CBA19D	7.00E-12	23% (48/201)	1			185

BM-005-B21.gene7	BM-005-B21.gene7		PREDICTED: similar to transposase n=2 Tax=Strongylocentrotus purpuratus RepID=UPI0000E4A2AF	1.00E-61	42% (110/259)	1	Replication, recombination and repair		614
BM-005-B21.gene8			unclassified						
BM-005-B21.gene9			unclassified						
BM-005-G14.gene1	BM-005-G14.gene1		papilin, putative n=1 Tax=Pediculus humanus corporis RepID=UPI000186ED45	0	49% (311/628)	2			1692
BM-005-G14.gene2	BM-005-G14.gene2		PREDICTED: similar to pogo transposable element with KRAB domain n=1 Tax=Hydra magnipapillata RepID=UPI0001925994	5.00E-23	36% (74/203)	1			276
BM-005-G14.gene3	BM-005-G14.gene3		papilin, putative n=1 Tax=Pediculus humanus corporis RepID=UPI000186ED45	0	34% (448/1316)	1			2054
BM-005-G14.gene4			unclassified						
BM-005-G14.gene5	BM-005-G14.gene5		Serpin 2, putative n=1 Tax=Ixodes scapularis RepID=B7PKZ6_IXOSC	3.00E-33	41% (90/217)	1			375
BM-006-B07.gene1	BM-006-B07.gene1		Transpanin, putative n=1 Tax=Ixodes scapularis RepID=B7QNZ7_IXOSC	4.00E-15	59% (39/66)	1			207
BM-006-B07.gene10	BM-006-B07.gene10		PREDICTED: similar to LReO_3 n=1 Tax=Danio rerio RepID=UPI0001760A86	1.00E-118	34% (257/752)	2			1113
BM-006-B07.gene11	BM-006-B07.gene11		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	5.00E-19	32% (53/161)	1			240
BM-006-B07.gene12			unclassified						
BM-006-B07.gene2	BM-006-B07.gene2		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	3.00E-26	31% (54/173)	1			241
BM-006-B07.gene3	BM-006-B07.gene3		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	3.00E-62	33% (158/474)	1			622
BM-006-B07.gene4			unclassified						
BM-006-B07.gene5			unclassified						
BM-006-B07.gene6			unclassified						
BM-006-B07.gene7			unclassified						
BM-006-B07.gene8	BM-006-B07.gene8		Endonuclease-reverse transcriptase n=1 Tax=Schistosoma mansoni RepID=Q4QQE2_SCHMA	3.00E-13	31% (45/143)	1			196
BM-006-B07.gene9	BM-006-B07.gene9		rve multi-domain protein	8.00E-10	22% (26/114)	1			143
BM-010-J12.gene1			unclassified						
BM-010-J12.gene10	BM-010-J12.gene10		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	8.00E-13	30% (60/198)	1			190
BM-010-J12.gene11	BM-010-J12.gene11		PREDICTED: similar to endonuclease and reverse transcriptase-like protein n=1 Tax=Tribolium castaneum RepID=UPI0000D578A9	3.00E-06	34% (32/93)	4			139
BM-010-J12.gene12			unclassified						
BM-010-J12.gene13			unclassified						
BM-010-J12.gene14	BM-010-J12.gene14		Gag-like protein n=1 Tax=Monascus pilosus RepID=Q2PWB5_9EURO	1.00E-06	35% (28/79)	7			142
BM-010-J12.gene15	BM-010-J12.gene15		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	1.00E-11	29% (58/198)	1			179

BM-010-J12.gene16	BM-010-J12.gene16		Reverse transcriptase, putative n=1 Tax=Ixodes scapularis RepID=B7PS32_IXOSC	1.00E-05	37% (32/86)	2			130
BM-010-J12.gene17			unclassified						
BM-010-J12.gene18			unclassified						
BM-010-J12.gene19			unclassified						
BM-010-J12.gene2	BM-010-J12.gene2		PREDICTED: similar to endonuclease and reverse transcriptase-like protein n=1 Tax=Tribolium castaneum RepID=UPI0000D578A9	1.00E-06	34% (32/93)	4			141
BM-010-J12.gene20			unclassified						
BM-010-J12.gene21	BM-010-J12.gene21		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	2.00E-55	38% (115/295)	1			563
BM-010-J12.gene22	BM-010-J12.gene22		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	4.00E-75	38% (148/382)	1			731
BM-010-J12.gene23	BM-010-J12.gene23		Putative gag protein n=1 Tax=Philodina roseola RepID=A1YGR8_9BILA	4.00E-13	30% (47/156)	1			195
BM-010-J12.gene24	BM-010-J12.gene24		Gap-Pol polyprotein n=1 Tax=Schistosoma japonicum RepID=C7C1Z1_SCHJA	6.00E-06	36% (25/68)	2			126
BM-010-J12.gene3	BM-010-J12.gene3		Predicted protein n=1 Tax=Nematostella vectensis RepID=A7S4I1_NEMVE	2.00E-11	33% (47/141)	2			182
BM-010-J12.gene4	BM-010-J12.gene4	yfeT	HTH-type transcriptional regulator murR n=49 Tax=Enterobacteriaceae RepID=MURR_ECO24	1.00E-111	99% (206/208)	1	Transcription		1004
BM-010-J12.gene5	BM-010-J12.gene5		Polyprotein of viral origin, putative (Fragment) n=1 Tax=Ixodes scapularis RepID=B7QCW6_IXOSC	8.00E-76	55% (136/246)	1			743
BM-010-J12.gene6			unclassified						
BM-010-J12.gene7	BM-010-J12.gene7		Pol protein n=1 Tax=Drosophila melanogaster RepID=Q9GP60_DROME	2.00E-35	21% (145/665)	1			390
BM-010-J12.gene8	BM-010-J12.gene8		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	8.00E-54	32% (120/370)	1			538
BM-010-J12.gene9	BM-010-J12.gene9		PREDICTED: similar to GA18855-PA n=1 Tax=Tribolium castaneum RepID=UPI0000D56A7F	3.00E-23	40% (59/144)	1			286
BM-012-E08.gene1			unclassified						
BM-012-E08.gene2	BM-012-E08.gene2		SSU rRNA; Hyalomma lusitanicum	2.00E-33	98% (76/77)	1			73
BM-012-E08.gene3	BM-012-E08.gene3		SSU rRNA; Hyalomma rufipes	2.00E-67	100% (131/131)	1			131
BM-013-M17.gene1	BM-013-M17.gene1		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	1.00E-36	36% (93/253)	1			401
BM-013-M17.gene2			unclassified						
BM-013-M17.gene3	BM-013-M17.gene3		Unassigned protein	5.00E-36	48% (75/155)	0			390
BM-013-M17.gene4			unclassified						
BM-013-M17.gene5	BM-013-M17.gene5		Endonuclease-reverse transcriptase n=1 Tax=Schistosoma japonicum RepID=C7C208_SCHJA	4.00E-19	31% (73/233)	2			249
BM-013-M17.gene6	BM-013-M17.gene6		Reverse transcriptase, putative n=1 Tax=Ixodes scapularis RepID=B7PS32_IXOSC	1.00E-16	36% (56/152)	8			230
BM-013-M17.gene7	BM-013-M17.gene7		Endonuclease-reverse transcriptase n=1 Tax=Schistosoma mansoni	6.00E-09	29% (33/112)	1			159

			RepID=Q4QQE0_SCHMA						
BM-013-M17.gene8	BM-013-M17.gene8		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A3B	1.00E-18	37% (50/132)	1			209
BM-013-M17.gene9			unclassified						
BM-026-P08.gene1	BM-026-P08.gene1		AGAP005534-PA n=1 Tax=Anopheles gambiae RepID=Q5TRL1_ANOGA	4.00E-06	24% (51/208)	6			130
BM-026-P08.gene10	BM-026-P08.gene10		Tymo_45kd_70kd domain containing protein	1.00E-10	22% (75/327)	1			148
BM-026-P08.gene11	BM-026-P08.gene11		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	4.00E-15	32% (56/171)	1			214
BM-026-P08.gene2			unclassified						
BM-026-P08.gene3	BM-026-P08.gene3		PREDICTED: similar to polyprotein n=1 Tax=Danio rerio RepID=UPI0001760EBD	5.00E-11	36% (24/65)	2			123
BM-026-P08.gene4			unclassified						
BM-026-P08.gene5	BM-026-P08.gene5		Asp_protease multi-domain protein	1.00E-05	31% (30/96)	1			108
BM-026-P08.gene6	BM-026-P08.gene6		Polyprotein of viral origin, putative (Fragment) n=1 Tax=Ixodes scapularis RepID=B7QCW6_IXOSC	1.00E-77	56% (139/246)	1			760
BM-026-P08.gene7	BM-026-P08.gene7		Reverse transcriptase, putative n=1 Tax=Ixodes scapularis RepID=B7PTE6_IXOSC	7.00E-29	44% (69/154)	1			332
BM-026-P08.gene8	BM-026-P08.gene8		PREDICTED: similar to GA18855-PA n=1 Tax=Tribolium castaneum RepID=UPI0000D56A7F	2.00E-34	38% (82/215)	1			381
BM-026-P08.gene9	BM-026-P08.gene9		Putative pol protein n=4 Tax=Adineta vaga RepID=A1YGS2_ADIVA	9.00E-09	24% (66/269)	1			160
BM-031-L02.gene1			unclassified						
BM-031-L02.gene10			unclassified						
BM-031-L02.gene11			unclassified						
BM-031-L02.gene12			unclassified						
BM-031-L02.gene13	BM-031-L02.gene13		predicted protein n=1 Tax=Pediculus humanus corporis RepID=UPI000186EC16	4.00E-05	31% (33/105)	2			125
BM-031-L02.gene14			unclassified						
BM-031-L02.gene2	BM-031-L02.gene2		PREDICTED: similar to protein F28E10.3 [imported] - Caenorhabditis elegans n=2 Tax=Strongylocentrotus purpuratus RepID=UPI0000E49859	2.00E-16	30% (49/160)	1			221
BM-031-L02.gene3	BM-031-L02.gene3		PREDICTED: similar to GA18855-PA n=1 Tax=Tribolium castaneum RepID=UPI0000D56A7F	1.00E-32	34% (90/262)	1			370
BM-031-L02.gene4			unclassified						
BM-031-L02.gene5			unclassified						
BM-031-L02.gene6			unclassified						
BM-031-L02.gene7	BM-031-L02.gene7		Reverse transcriptase, putative n=1 Tax=Ixodes scapularis RepID=B7PS32_IXOSC	3.00E-15	34% (51/147)	6			220
BM-031-L02.gene8			unclassified						
BM-031-L02.gene9			unclassified						
BM-066-M07.gene1	BM-066-M07.gene1		GL14012 n=1 Tax=Drosophila persimilis RepID=B4GNG1_DROPE	0	34% (416/1203)	1			1710
BM-066-M07.gene10			unclassified						

BM-066-M07.gene11			unclassified						
BM-066-M07.gene12	BM-066-M07.gene12		ATP-dependent RNA helicase, putative n=1 Tax=Ixodes scapularis RepID=B7PA01_IXOSC	2.00E-21	38% (59/152)	1			263
BM-066-M07.gene13	BM-066-M07.gene13		Cytochrome P450 n=2 Tax=Rhipicephalus microplus RepID=D4PH83_BOOMI	4.00E-22	89% (51/57)	1	Secondary metabolites biosynthesis, transport and catabolism		272
BM-066-M07.gene14	BM-066-M07.gene14		Cytochrome P450 n=2 Tax=Rhipicephalus microplus RepID=D4PH83_BOOMI	7.00E-91	78% (173/219)	1	Secondary metabolites biosynthesis, transport and catabolism		868
BM-066-M07.gene15	BM-066-M07.gene15		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	8.00E-17	40% (53/130)	1			225
BM-066-M07.gene16	BM-066-M07.gene16		Putative gag protein n=1 Tax=Philodina roseola RepID=A1YGR8_9BILA	5.00E-09	25% (41/158)	2			163
BM-066-M07.gene2			unclassified						
BM-066-M07.gene3			unclassified						
BM-066-M07.gene4			unclassified						
BM-066-M07.gene5			unclassified						
BM-066-M07.gene6			unclassified						
BM-066-M07.gene7	BM-066-M07.gene7		Unassigned protein	8.00E-06	30% (22/73)	0			91
BM-066-M07.gene8			unclassified						
BM-066-M07.gene9	BM-066-M07.gene9	ddx5	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5 n=4 Tax=Xenopus RepID=Q6P4W2_XENTR	6.00E-91	42% (169/396)	1		Spliceosome	828
BM-074-Random-F12.gene1			unclassified						
BM-074-Random-F12.gene10			unclassified						
BM-074-Random-F12.gene11			unclassified						
BM-074-Random-F12.gene12			unclassified						
BM-074-Random-F12.gene13			unclassified						
BM-074-Random-F12.gene14			unclassified						
BM-074-Random-F12.gene15	BM-074-Random-F12.gene15		Ferrichrome-binding protein (Fragment) n=1 Tax=Bacillus cereus W RepID=B3Z1U9_BACCE	2.00E-34	77% (73/94)	2			372
BM-074-Random-F12.gene2	BM-074-Random-F12.gene2		PREDICTED: similar to gag-pol polyprotein n=2 Tax=Strongylocentrotus purpuratus RepID=UPI0000E4A4EC	4.00E-40	38% (109/282)	1			427
BM-074-Random-F12.gene3			unclassified						
BM-074-Random-F12.gene4			unclassified						
BM-074-Random-F12.gene5			unclassified						
BM-074-Random-F12.gene6			unclassified						
BM-074-Random-F12.gene7			unclassified						
BM-074-Random-F12.gene8			unclassified						
BM-074-Random-F12.gene9			unclassified						
BM-077-G20.gene1	BM-077-G20.gene1		Cytochrome P450 n=2 Tax=Rhipicephalus microplus RepID=D4PH83_BOOMI	0	68% (352/512)	1	Secondary metabolites biosynthesis, transport and catabolism		1799
BM-077-G20.gene2			unclassified						
BM-077-G20.gene3	BM-077-G20.gene3		Cytochrome P450 n=2 Tax=Rhipicephalus microplus RepID=D4PH83_BOOMI	1.00E-108	83% (213/256)	1	Secondary metabolites biosynthesis, transport and		1021

							catabolism		
BM-077-G20.gene4			unclassified						
BM-077-G20.gene5			unclassified						
BM-077-G20.gene6	BM-077-G20.gene6		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	2.00E-73	38% (148/381)	1			720
BM-077-G20.gene7	BM-077-G20.gene7		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	8.00E-17	40% (53/130)	1			225
BM-077-G20.gene8	BM-077-G20.gene8		Cytochrome P450 n=2 Tax=Rhipicephalus microplus RepID=D4PH83_BOOMI	1.00E-90	79% (170/214)	1	Secondary metabolites biosynthesis, transport and catabolism		865
BM-077-Random-J09.gene1	BM-077-Random-J09.gene1		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	2.00E-19	43% (64/147)	1			254
BM-077-Random-J09.gene10	BM-077-Random-J09.gene10		Unassigned protein	7.00E-18	41% (46/110)	0			239
BM-077-Random-J09.gene11			unclassified						
BM-077-Random-J09.gene12			unclassified						
BM-077-Random-J09.gene2	BM-077-Random-J09.gene2		Pol-like protein n=1 Tax=Biomphalaria glabrata RepID=A3FMR3_BIOGL	1.00E-35	33% (86/260)	1			390
BM-077-Random-J09.gene3	BM-077-Random-J09.gene3		Gag-like protein n=1 Tax=Biomphalaria glabrata RepID=A3FMR2_BIOGL	1.00E-35	34% (80/234)	1			388
BM-077-Random-J09.gene4	BM-077-Random-J09.gene4		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	3.00E-63	36% (144/391)	1			631
BM-077-Random-J09.gene5	BM-077-Random-J09.gene5		PREDICTED: similar to pol polyprotein n=1 Tax=Tribolium castaneum RepID=UPI000175895B	2.00E-76	42% (156/366)	1			747
BM-077-Random-J09.gene6			unclassified						
BM-077-Random-J09.gene7			unclassified						
BM-077-Random-J09.gene8			unclassified						
BM-077-Random-J09.gene9			unclassified						
BM-118-H10.gene1			unclassified						
BM-118-H10.gene10			unclassified						
BM-118-H10.gene2	BM-118-H10.gene2		Unassigned protein	5.00E-09	35% (33/92)	0			156
BM-118-H10.gene3	BM-118-H10.gene3		PREDICTED: similar to pol polyprotein n=1 Tax=Danio rerio RepID=UPI0001761336	4.00E-21	50% (49/98)	1			270
BM-118-H10.gene4	BM-118-H10.gene4		Pol-like protein n=1 Tax=Biomphalaria glabrata RepID=A3FMR3_BIOGL	8.00E-14	31% (51/161)	1			202
BM-118-H10.gene5	BM-118-H10.gene5		Gag-like protein n=1 Tax=Biomphalaria glabrata RepID=A3FMR2_BIOGL	3.00E-09	32% (50/154)	2			161
BM-118-H10.gene6	BM-118-H10.gene6		PREDICTED: similar to GA18855-PA n=1 Tax=Tribolium castaneum RepID=UPI0000D56A7F	8.00E-40	35% (95/267)	1			430
BM-118-H10.gene7			unclassified						
BM-118-H10.gene8	BM-118-H10.gene8		OSJNBa0088I22.10 protein n=1 Tax=Oryza sativa Japonica Group RepID=Q7XU43_ORYSJ	3.00E-15	37% (37/99)	1			213
BM-118-H10.gene9	BM-118-H10.gene9		PREDICTED: similar to pol-like protein n=1 Tax=Nasonia vitripennis RepID=UPI00015B43EC	1.00E-09	36% (40/110)	1			169
BM-129-CzEst9-	BM-129-CzEst9-		Predicted protein n=1 Tax=Nematostella	1.00E-18	32% (46/142)	1			220

N14.gene1	N14.gene1		vectensis RepID=A7T484_NEMVE						
BM-129-CzEst9-N14.gene10			unclassified						
BM-129-CzEst9-N14.gene11	BM-129-CzEst9-N14.gene11		PREDICTED: similar to pol polyprotein n=1 Tax=Danio rerio RepID=UPI0001761336	5.00E-20	51% (46/90)	1			261
BM-129-CzEst9-N14.gene12			unclassified						
BM-129-CzEst9-N14.gene13	BM-129-CzEst9-N14.gene13		Unassigned protein						
BM-129-CzEst9-N14.gene14	BM-129-CzEst9-N14.gene14		PREDICTED: similar to protease, reverse transcriptase, ribonuclease H, integrase n=2 Tax=Tribolium castaneum RepID=UPI000175894A	1.00E-16	32% (56/171)	1			223
BM-129-CzEst9-N14.gene2			unclassified						
BM-129-CzEst9-N14.gene3	BM-129-CzEst9-N14.gene3		PREDICTED: similar to polyprotein n=1 Tax=Tribolium castaneum RepID=UPI0001758A13	4.00E-81	36% (175/473)	1			783
BM-129-CzEst9-N14.gene4			unclassified						
BM-129-CzEst9-N14.gene5	BM-129-CzEst9-N14.gene5		PREDICTED: similar to zinc finger protein, partial n=1 Tax=Strongylocentrotus purpuratus RepID=UPI0000E48E5A	2.00E-88	31% (233/737)	1			852
BM-129-CzEst9-N14.gene6			unclassified						
BM-129-CzEst9-N14.gene7			unclassified						
BM-129-CzEst9-N14.gene8			unclassified						
BM-129-CzEst9-N14.gene9	BM-129-CzEst9-N14.gene9		Acetylcholinesterase, putative n=1 Tax=Ixodes scapularis RepID=B7P512_IXOSC	4.00E-84	33% (188/557)	8			810

Appendix 5.2. Primer sequences and positions in *R. microplus* BAC sequences, Table 1 BAC BM-012-E08 primers, Table 2 and 3 BM-005-G14 papilin and helicase genes, Table 4 BM-005-G14 serpin gene.

Table 1

rRNA	5' postion	forward_sequence	3' position	reverse_sequence	Contigs	
15K	12076	TAGTGACGCGCATGAATGG	12367	CCATGGGACATCAACAACC	BM012E08_c2	
17K.0		TACGAGGCTGAGATCATTGC		TCTTTGCTCGCTTCTTTTGC		
18K	351	ACATTCCTTCCCTCCCTTCC	623	GCTCAACAGGGTCTTCTTTCC	BM012E08_c1	
20K	1689	GCTTAGCCTCTGACTGGAAGG	1871	TGGTTCCTTTCGTACTTCACG	BM012E08_c1	
Repeat						
22K.1	2817	AGTTCGCTTGCTCTGACACC	3092	GCCAAGTACGAAAGGAACAGC	BM012E08_c1	
22K.2	2994	TCTCCGACACCCATATTTCG	3232	AGAGGTGCCTGAGAGAAACG	BM012E08_c1	
17K.1	358	TGGATGTACTCAAGGGTACGG	658	GCTGAAGTGATTTTTCCTGTCC	BM012E08_c4	
17K.2	190	AACAAGTGGCAACAACAAGC	475	CACAGTCGCCTTTTGTGAGG	BM012E08_c4	
38K.1	508	TTTTCCTTAACGACGTCTCG	793	GTATTTGGACGTGCTTGACG	BM012E08_c3	
38K.2	3207	GTTTTCCTTAACGACGTCTCG	3493	GTATTTGGACGTGCTTGACG	BM012E08_c2	
rRNA						
6K	9143	ATTCAGCGGGTTGTCTCG	9359	ATGGGTTTACGAACGTGTCC	BM012E08_c2	
3K	11980	GGACAATATTCTACGGAACAGG	12159	ATGTGATTTCTGCCCAGTGC	BM012E08_c2	
9K		AGAAATCACATTGCGTCAGC	11971	GCCTTCCTGGGGAAAAGC	BM012E08_c2	
8K	1538	GAACGTTTCGGTTTCGTACC	1738	GGAGGCCGAGTATTGACG	BM012E08_c1	
long range primer					**Fwd**	**Rev**
rDNA	4585	GTGGTGGCTGCTCTACTACTTACG	1672	TTAAATCAGTTATGGTTCCTTAGATCG	BM012E08_c2	BM012E08_c1
rDNA	4611	CTACTACTTACGACACCACGACAGG	1660	GTTTCTTCCTACTTGGATAACTGTGG	BM012E08_c2	BM012E08_c1
Intergenic region	1689	AAGTAGGAAGAAACGATCTAAGGAACC	4624	GCTTAGCCTCTGACTGGAAGG	BM012E08_c1	BM012E08_c2
Intergenic region	1663	TTAGCTCTAGAATTGCCACAGTTATCC	4651	GTCGTGGTGTCGTAAGTAGTAGAGC	BM012E08_c1	BM012E08_c2

Table 2

Primers	sequence 5' - 3'	5' position
PapStartF2	ATGAGGAGTAGGGGAGCTGTGCTCCT	175
Papilin54900R	CCATTGACGCAAACATCAAG	728
Papilin57383F	TGTGACAAGATGCTGGGTTC	748
AdamSR1	CGATGGGATCAGCAGGATGTCGTT	891
3425F	GACAAGGTTGAAGTTCTGGATGA	1651
Pap5360F	CTGCTGTCCAGATGGCATCAC	2619
Pap7260F	CGAAGGATGCACCTGTGAGCA	3453
Pap8020F	TACTCCGGCTCGTGGACCTAA	4203
Pap8760R	GCCTCGCACTCTTCCTCCGA	4805
Pap9690R	GCACATCTGTTGGCACTCCAG	5703
Pap9780R	CTGCAGCTCTTCGCAGTTGGA	6744
pap12440F	CGAAACAGACGGTCATCGCGTT	7950
PapilinR3	CATTTCCATTACTGTACCCATTGACGTA	8470
PapStopR1	TTGCAGATTTGGCAAGATTCGCTTAGGC	8560
pap13230R	TAATGTATTGGTATTAGACTATGGGT	8761
papRT-F6	CCTGATGGCCGTACGGCAGCTA	4417
papRT-R6	CCTCGAGCTGGAGTTCGCCCAT	4517

Table 3

Primers	sequence 5' - 3'	5' position
PapilinORFF2	ATGGAAGGCAGTGGTGATCT	1
PapRT-F	CCGACGCCTGGAGCTGTTA	124
Pap1090F	CGAGTCGAAGGGACCAGTGTA	976
Pap2440R	GTGACACTTCAATACTTGCCACT	2900
Pap2110F	AGCACTCGTTTCAGACCGTCT	1981
Pap2350R	ACACGCATAGACATGCACTAGT	3892
papilin50819R	TGTAGTAAGGCACCCAGCTGTA	4759
Papilin54900R*	CCATTGACGCAAACATCAAG	4886
papRT-F3	ACCAGTGTACTCCTGTCGGCA	988
papRT-R3	CTACGTCAACGGTCATTGGATCT	1053

Table 4

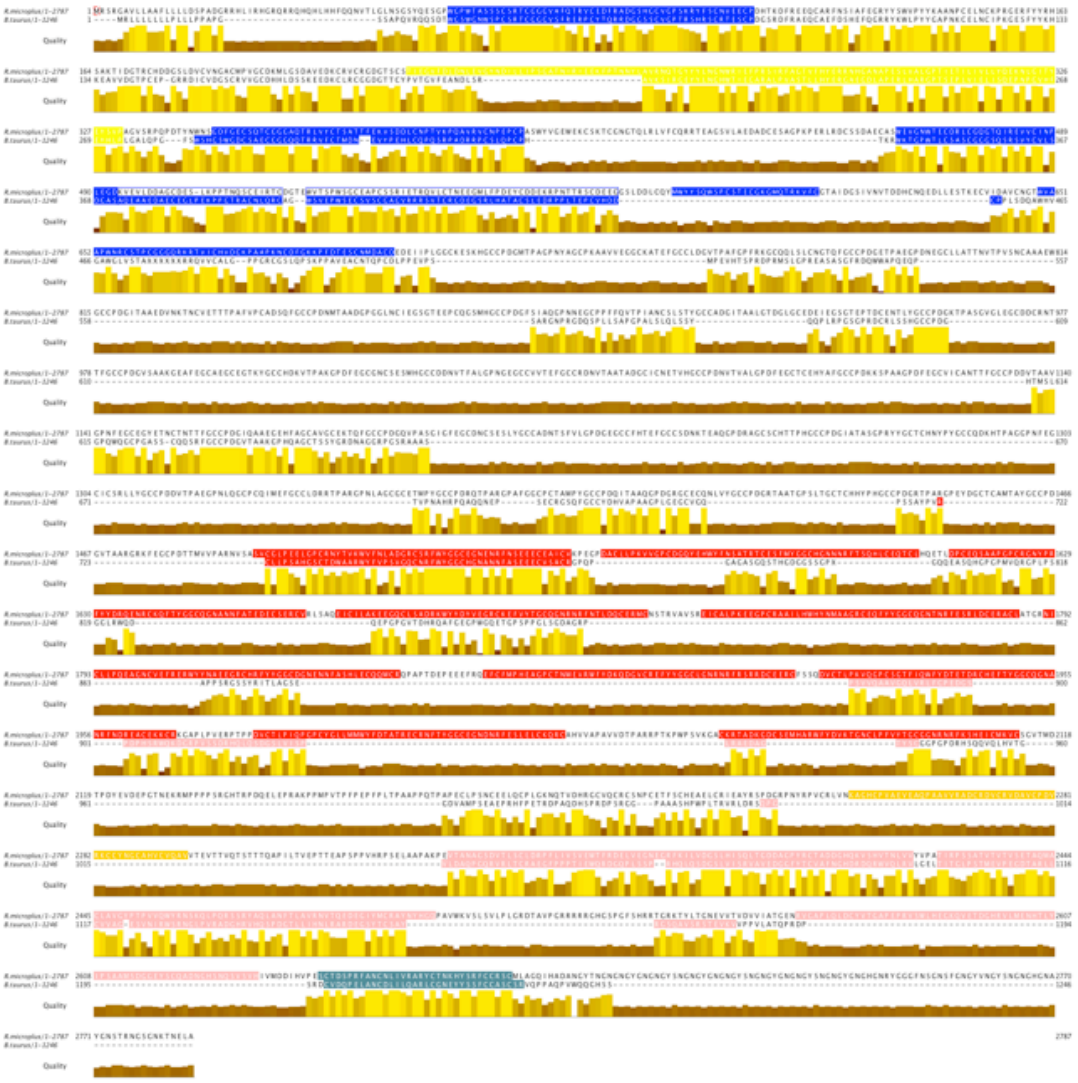
Label	Primers	BAC position(bp)
SerpF1	AAACGCGTGGCGATTGCCGACA	135040
SerpF2	GTAGAAGATACGACAGATCTGAAGG	133688
SerpF3	TCCTCTCGTAAGGTACGAGTGGA	129586
SerpR1	CAAAGGTTGTTGTCACCTTGTGCG	123298
SerpR2	GAAATCTTCGTCCCGTATGTTCTG	125843
SerpR3	ACTGCGGCAACCCAGCGTCTC	129470

Appendix 5.3 BES analysis to *Bmi*GI version 2.1

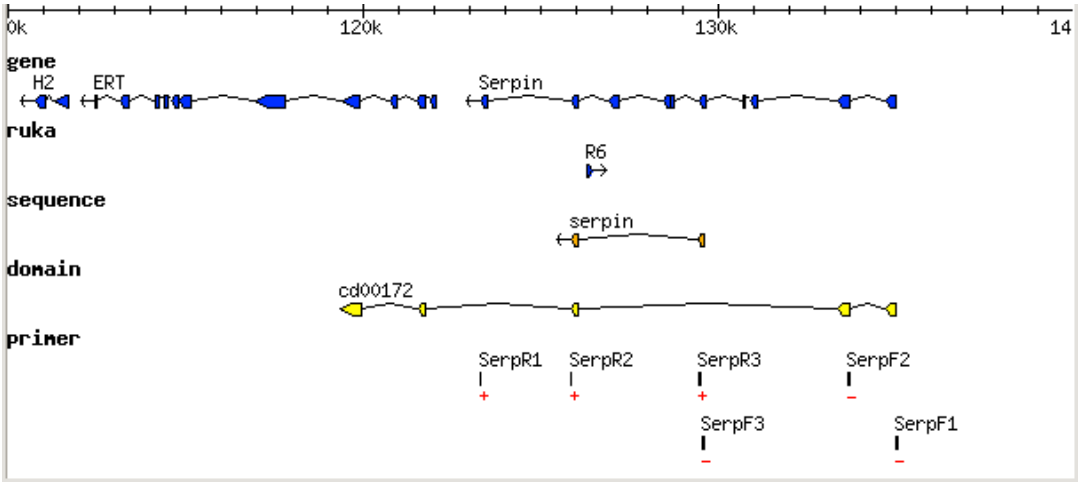
DFCI BmiGI	Query accession	Query length	Start ('query')	End ('query')	Accession	hit_length	Start ('hit')	End ('hit')	Description	evalue	Percent identity	Hsp length (aa)	Coverage query	Coverage hit	Strand ('query')	Strand ('hit')
	3-BM-005_G14_F.trimmed	771	590	615	CV446812	823	528	553	weakly similar to UniRef100_Q7Q6M0 Cluster: AGAP005771-PA; n=1; Anopheles gambiae str. PEST|Rep: AGAP005771-PA - Anopheles gambiae str. PEST, partial (4%)	5.00E-04	96.15384615	26	3.242542153	3.037667072	1	-1
	3-BM-005_G14_F.trimmed	771	1	18	TC24112	241	191	208	similar to UniRef100_P0C6A2 Cluster: Mastermind-like domain-containing protein 1; n=1; Mus musculus|Rep: Mastermind-like domain-containing protein 1 - Mus musculus (Mouse), partial (3%)	0.13	100	18	2.33463035	7.468879668	1	1
	3-BM-005_G14_R.trimmed	734	122	144	CV453292	449	395	417	similar to UniRef100_A6DUI4 Cluster: Sodium:dicarboxylate symporter family/bacterial extracellular solute- binding family 3, partial (2%)	1.00E-04	100	23	3.133514986	5.122494432	1	-1
	3-BM-005_G14_R.trimmed	734	122	144	TC21417	1191	842	864	similar to UniRef100_A7PT99 Cluster: Chromosome chr8 scaffold_29, whole genome shotgun sequence; n=1; Vitis vinifera|Rep: Chromosome chr8 scaffold_29, whole genome shotgun sequence - Vitis vinifera (Grape), partial (4%)	1.00E-04	100	23	3.133514986	1.931150294	1	1
	3-BM-012_E08_F.trimmed	734	555	734	TC20442	757	1	180	gb|AF018654.1|AF018654 Rhipicephalus zambesiensis B304 18S ribosomal RNA gene, partial sequence, partial (29%)	3.00E-98	100	180	24.52316076	23.77807133	1	1
	3-BM-012_E08_F.trimmed	734	146	276	CV440198	131	1	131	gb|AF018653.1|AF018653 Rhipicephalus appendiculatus B301 18S ribosomal RNA gene, partial sequence, partial (7%)	5.00E-69	100	131	17.847411 44	100	1	1
	3-BM-012_E08_R.trimmed	43	1	38	TC17450	1865	1282	1319	similar to UniRef100_Q8UWJ5 Cluster: CDH1-D; n=1; Gallus gallus|Rep: CDH1-D - Gallus gallus (Chicken), partial (26%)	6.00E-15	100	38	88.37209302	2.037533512	1	-1
	3-BM-	43	5	30	TC21524	751	590	614		1.2	92.3076923	26	55.8139534	3.1957390	1	1

	012_E08_R.trimmed										1		9	15		
NCBI CDD	**query_accession**	**Query length**	**Start ('query')**	**End ('query')**	**Accession**	**hit_ength**	**Start ('hit')**	**End ('hit')**	**Description**	**Significance**	**Percent identity**	**Hsp length (aa)**	**Coverage query**	**Coverage hit**	**Strand ('query')**	**Strand ('hit')**
	3-BM-005_G14_F.trimmed	771	235	387	32851	481	191	242	COG3037, SgaT, Uncharacterized protein conserved in bacteria [Function unknown].	1.1	25	52	5.058365759	2.702702703	1	0
	3-BM-005_G14_F.trimmed	771	61	153	71008	125	92	122	pfam07560, DUF1539, Domain of Unknown Function (DUF1539)..	1.7	25.80645161	31	3.112840467	6.4	1	0
	3-BM-005_G14_R.trimmed	734	306	629	29117	364	164	267	cd00172, SERPIN, SERine Proteinase INhibitors (serpins) exhibit conformational polymorphism shifting from native to cleaved, latent, delta, or polymorphic forms. Many serpins, such as antitrypsin and antichymotrypsin, function as serine protease inhibitors which regulate blood coagulation cascades. Non-inhibitory serpins perform many diverse functions such as chaperoning proteins or transporting hormones. Serpins are of medical interest because mutants have been associated with blood clotting disorders, emphysema, cirrhosis, and dementia..	1.00E-16	34.25925926	108	15.1226158	10.16483516	1	0
	3-BM-005_G14_R.trimmed	734	306	632	47438	360	159	262	smart00093, SERPIN, SERine Proteinase INhibitors; .	8.00E-12	33.02752294	109	14.71389646	10	1	0
	3-BM-012_E08_F.trimmed	734	328	612	84221	529	180	288	PRK12665, PRK12665, putative monovalent cation/H+ antiporter subunit D.	0.21	29.05982906	117	13.89645777	6.427221172	-1	0
	3-BM-012_E08_F.trimmed	734	132	212	65841	279	20	45	pfam02089, Palm_thioest, Palmitoyl protein thioesterase..	0.83	48.14814815	27	5.313351499	4.659498208	-1	0
SSH	**query_accession**	**Query length**	**Start ('query')**	**End ('query')**	**Accession**	**Hit length**	**Start ('hit')**	**End ('hit')**	**Description**	**Significance**	**Percent identity**	**Hsp length (aa)**	**Coverage query**	**Coverage hit**	**Strand ('query')**	**Strand ('hit')**
	3-BM-005_G14_F.trimmed	771	534	547	L3-3-A_Contig7	199	12	25	No definition line found [poly-A tail removed]	0.15	100	14	1.815823606	7.035175879	1	-1
	3-BM-005_G14_F.trimmed	771	534	547	F2-3-A_Clone_78	981	888	901	No definition line found [poly-T tail removed] [reverse complement]	0.15	100	14	1.815823606	1.427115189	1	-1
	3-BM-005_G14_R.trimmed	734	544	557	F1-2-A_Contig5	490	26	39	No definition line found	0.14	100	14	1.907356948	2.857142857	1	-1

	3-BM-005_G14_R.trimmed	734	704	715	M1-3-A_Contig1	741	514	525	No definition line found [poly-T tail removed] [reverse complement]	2.2	100	12	1.634877384	1.619433198	1	1
	3-BM-012_E08_F.trimmed	734	1	734	F2-3-A_Contig6	1439	160	898	No definition line found	0	97.29364005	739	97.95640327	49.96525365	1	1
	3-BM-012_E08_F.trimmed	734	173	734	L3-3-A_Contig4	1497	909	1470	No definition line found	0	99.28825623	562	76.02179837	37.2745491	1	-1
	3-BM-012_E08_R.trimmed	43	19	30	M1-2-A_Contig5	578	285	296	No definition line found [poly-A tail removed]	0.1	100	12	27.90697674	2.076124567	1	-1
	3-BM-012_E08_R.trimmed	43	29	39	F2-3-A_Clone_47	511	276	286	No definition line found [poly-T tail removed] [reverse complement]	0.41	100	11	25.58139535	2.152641879	1	1
BmiGlTC/ESTs	**query_accession**	**Query length**	**Start ('query')**	**End ('query')**	**Accession**	**Hit length**	**Start ('hit')**	**End ('hit')**	**Description**	**Significance**	**Percent identity**	**Hsp length (aa)**	**Coverage query**	**Coverage hit**	**Strand ('query')**	**Strand ('hit')**
	3-BM-005_G14_F.trimmed	771	590	615	BEAFB71TR	823	528	553		7.00E-04	96.15384615	26	3.242542153	3.037667072	1	-1
	3-BM-005_G14_F.trimmed	771	1	18	TC13887	241	191	208		0.18	100	18	2.33463035	7.468879668	1	1
	3-BM-005_G14_F.trimmed	771	533	550	TC9734	999	474	491		0.18	100	18	2.33463035	1.801801802	1	1
	3-BM-005_G14_R.trimmed	734	94	141	MPOA274TR	908	766	813		4.00E-05	87.5	48	5.722070845	4.625550661	1	-1
	3-BM-005_G14_R.trimmed	734	122	144	TC12565	1258	395	417		2.00E-04	100	23	3.133514986	1.828298887	1	-1
	3-BM-005_G14_R.trimmed	734	20	40	TC10842	1112	676	696		0.003	100	21	2.861035422	1.888489209	1	1
	3-BM-012_E08_F.trimmed	734	555	734	TC8779	757	1	180		4.00E-98	100	180	24.52316076	23.77807133	1	1
	3-BM-012_E08_F.trimmed	734	146	276	BEAE880TR	131	1	131		6.00E-69	100	131	17.84741144	100	1	1
	3-BM-012_E08_F.trimmed	734	313	458	BEAE880TF	146	1	146		1.00E-63	95.89041096	146	19.07356948	95.89041096	1	-1
	3-BM-012_E08_R.trimmed	43	1	38	MPOAL33TR	956	694	731		8.00E-15	100	38	88.37209302	3.974895397	1	-1
	3-BM-012_E08_R.trimmed	43	1	38	MPOAL33TF	934	547	584		8.00E-15	100	38	88.37209302	4.068522484	1	1
	3-BM-012_E08_R.trimmed	43	1	38	TC9245	1864	546	583		8.00E-15	100	38	88.37209302	2.038626609	1	1



Appendix 5.4. Papilin protein sequence alignment, *R. microplus* and *B. taurus*, domains are highlighted Kunitz BPTI (red), ADAM spacer1 (yellow), Ig-set (pink), PLAC (purple), WAP (orange), TSP1 (blue)

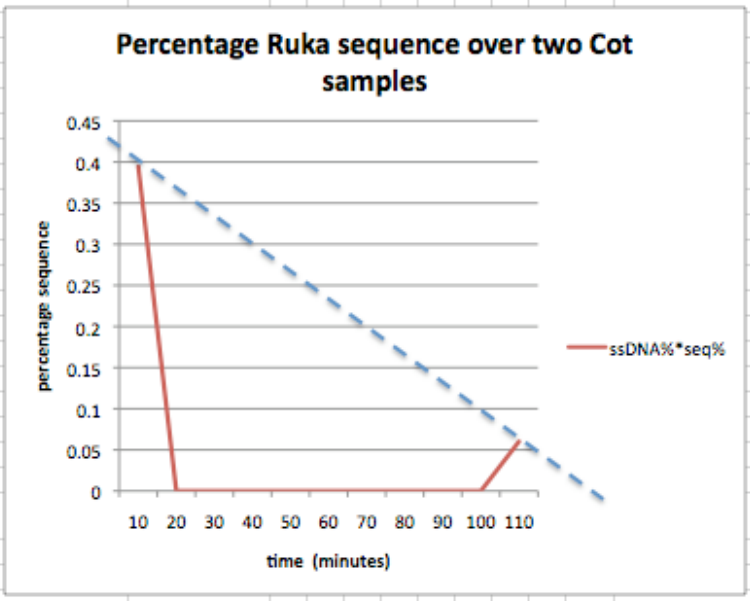


Appendix Figure 5.5 Serpin prediction in BAC BM-005-G14

Appendix 5.6. *R. microplus* Genome wide estimation of the Ruka element frequency based on two Cot re-association experiments.

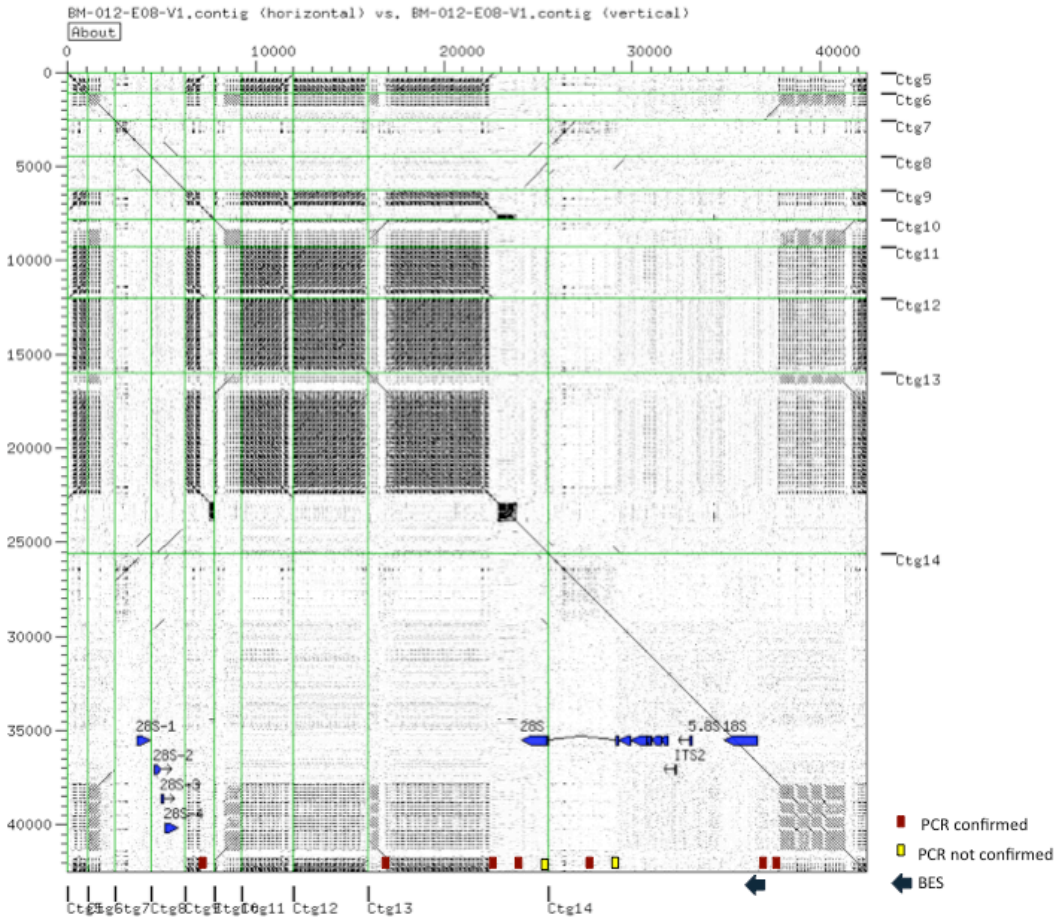
Label	**USDA696**	**USDA69**
Start [DNA] microgram	200	200
Renaturation time	1hr 48min 6sec	10min 49sec
Secs	6486	649
M.s	696	69
Cot	695.6	69.56
Renturation Temp	70	70
NaPO4 (M)	0.03	0.03
Ass260	0.954	
ssDNA (microgram)	2.52	18.8
dsDNA (microgram)	219.6	165
Total microgram DNA (ss & ds)	222.12	183.8
Fraction ssDNA	0.011345219	0.102285092
ssDNA %	1.13452188	10.22850925
Number reads	2,831,229	3,020,062
Reads (bp)	708,482,408	762,341,121
	8,037,888	77,976,132
Read length (bp)	250	252
G14 depth (bp)	869,692	1,063,349
G14 number reads	62,126	66,096
G14 length	135,000	135,000
G14 average coverage	13.99884106	16.08794783
	6.442162963	7.876659259
SINE Ruka	**USDA696**	**USDA69**
Total sequence read depth (bp)	46,048	40,997
Sequence length (bp)	195	195
Sequence average cov	236.1435897	210.2410256
Fraction sequence (%)	0.05294748	0.038554604
Sequence % * ssDNA%	0.060070075	0.394356127
Sequence micro = ssDNA microgram * Sequence %	0.13342765	0.724826562
Seq micro% = seq micro / Total microgram DNA (ss & ds)	0.000600701	0.003943561
% DNA at recovered time	0.060070075	0.394356127
K*secs=	0.999399299	0.996056439
K=	0.000154086	0.001534756
RUKA frequency	0.42%	
Genome size (bp)	7,100,000,000	
Ruka genome sequence (bp)	29,820,000	
Number RUKA elements	152,923	

	Cot69										Cot696
minutes	**10**	20	30	40	50	60	70	80	90	100	**110**
ssDNA%*seq%	0.394356127										0.060070075

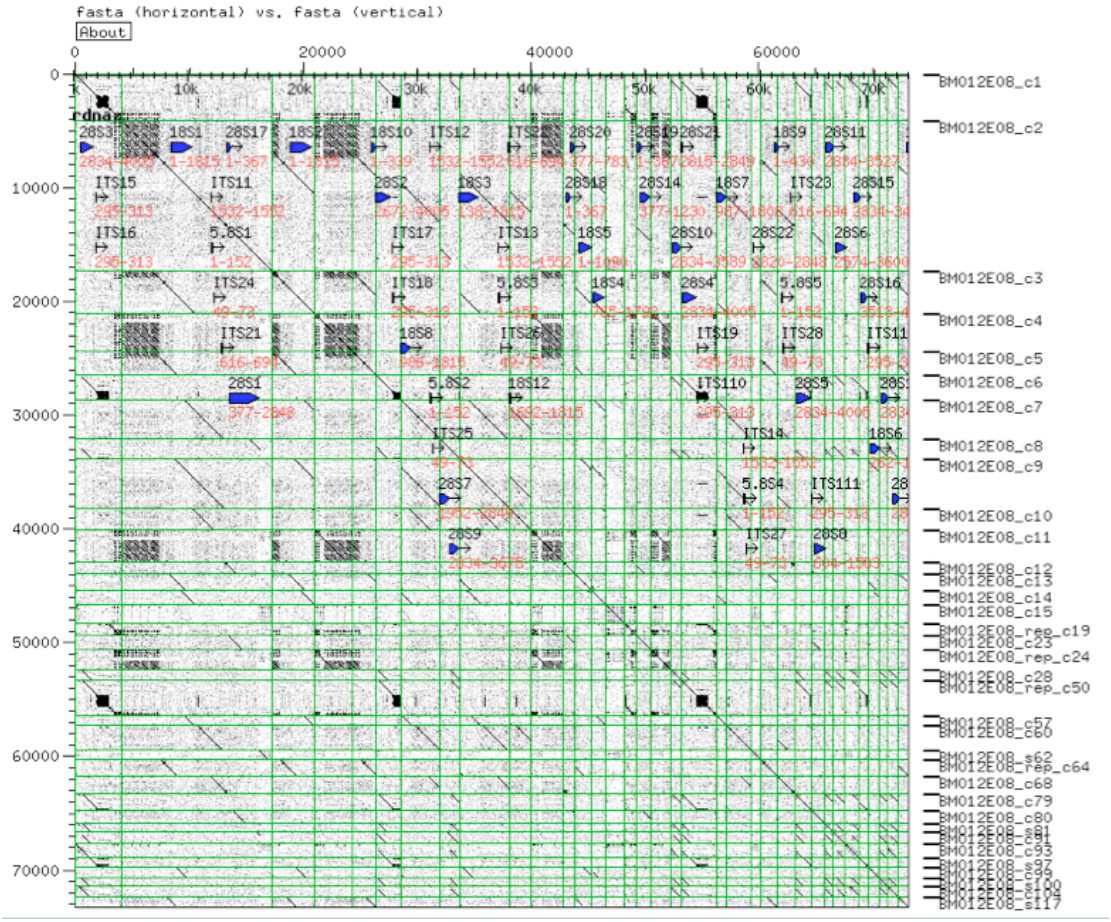


Appendix 5.7 Table of BAC sequence assembly statistics BM-012-E08

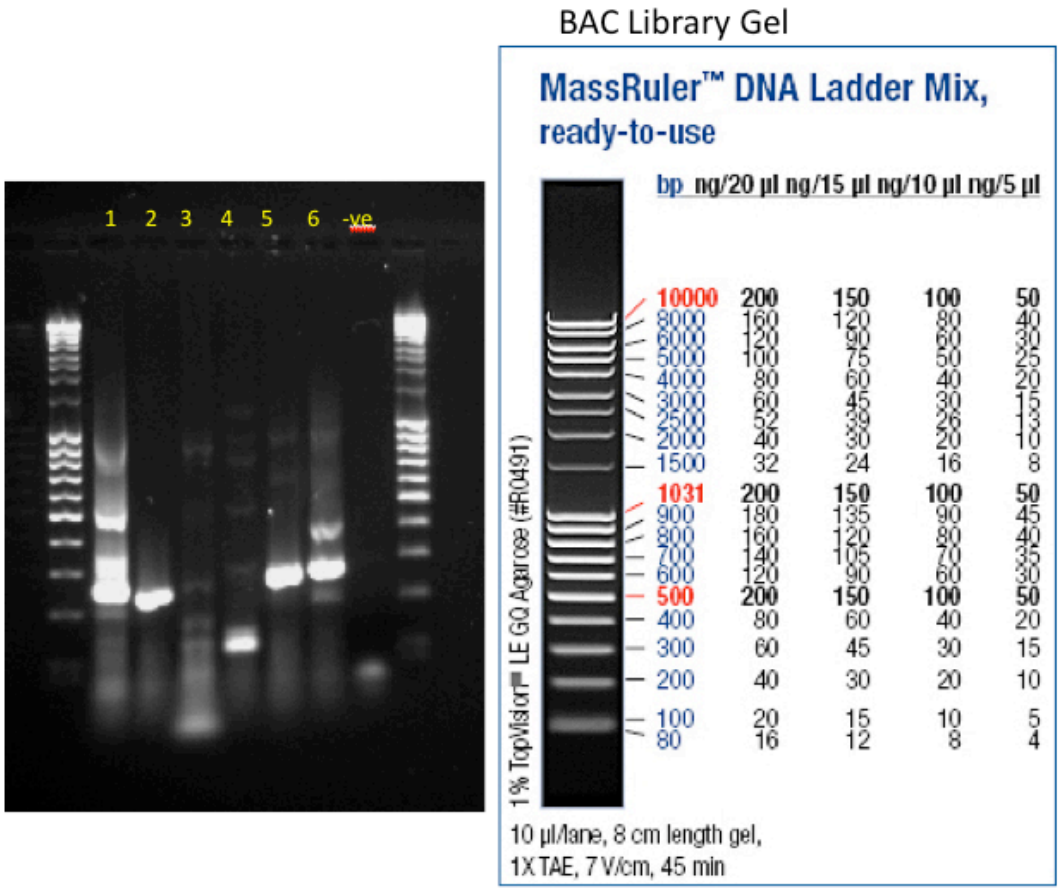
Mira Assembly	normal	repetitive
Num. reads assembled	1116	1102
Num. singlets	49	128
Large contigs		
Num. contigs	16	7
Total	44340	29969
Largest	10619	13221
N50	2685	4004
N90	1682	1774
N95	1393	1682
Coverage		
Max	47	46
Avg.	17.28	21.04
Quality		
Average	76	75
(WRMc):	16	9
All		
Num. contigs	70	49
Total	136767	169396
Largest	10619	13221
N50	1260	860
N90	700	590
N95	588	425
Coverage		
Max	47	46
Avg.	17.28	21.04
Quality		
Average	51	48
(WRMc):	16	14



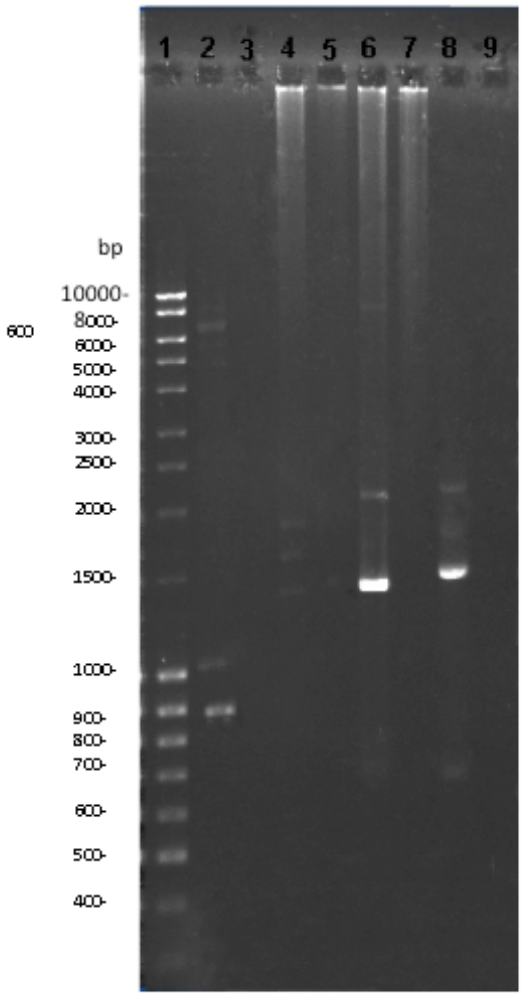
Appendix 5.8A. Dotmatrix of BAC BM-012-E08 non-redundant assembly reveals single unit Ctg11-14. Ambylomma rRNA (blue) fragments in Contigs 7,8,13,14. Lower horizontal axis shows regions confirmed (red), and those not confirmed by PCR (yellow), and the forward BES (green arrow)



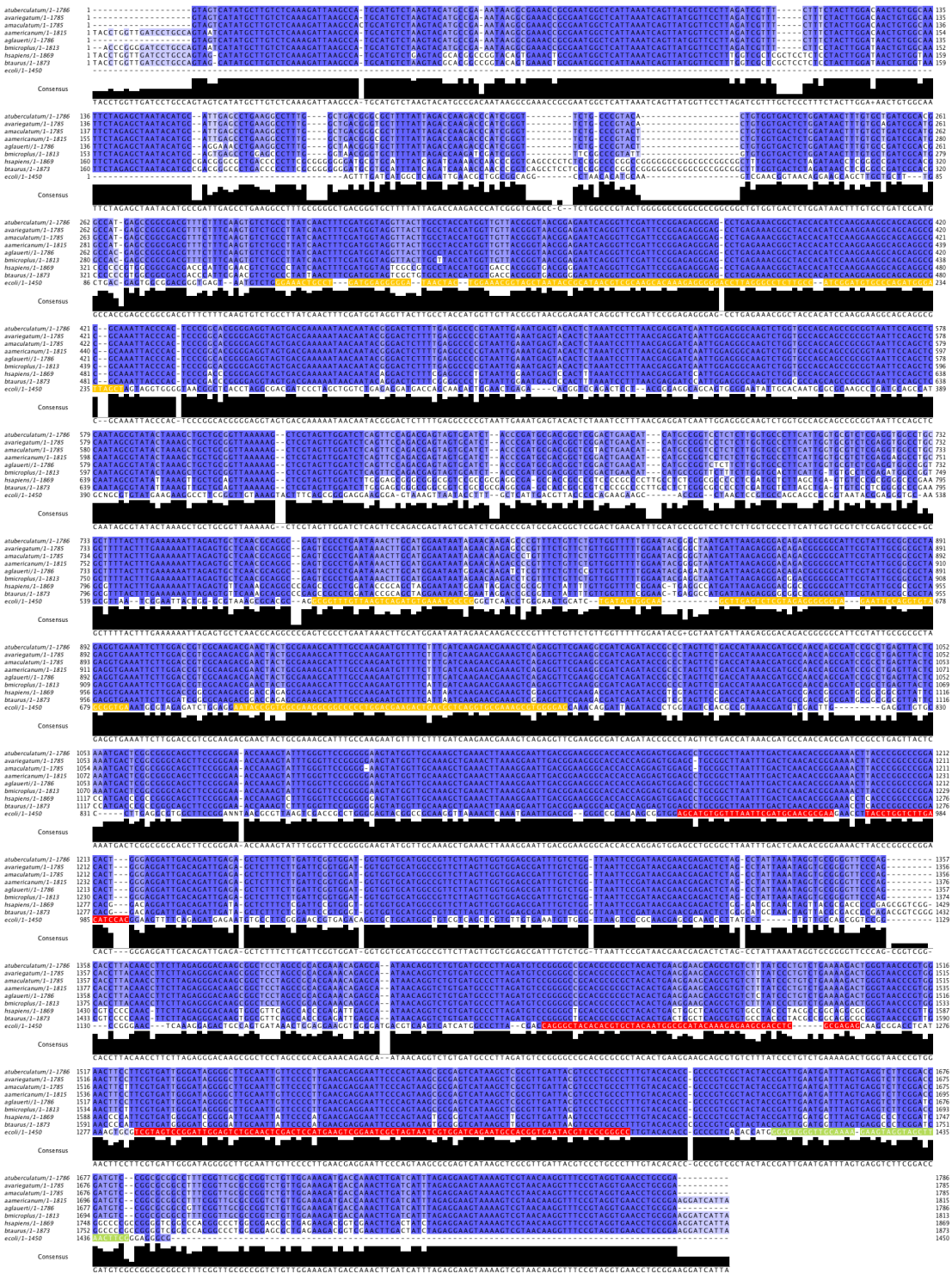
Appendix 5.8B. Dotmatrix of BAC BM-012-E08 redundant assembly reveals at least three units repeated, Ambylomma rRNA (blue).



Appendix 5.9A. PCR results for BM-012-E08 repetitive elements. Lanes: 1) 22K.1 2) 22K.2 3) 17K.1 4) 17K.2 5) 38K.1 6) 38K.2 7) -ve control



Appendix 5.9B. Long primer sets used to amplify tick genomic DNA using Roche Expand Long Template PCR system, Lane 1 Fermentas Mass ruler 80bp-10kb (#SM0403), Lane 2 rDNA.1, Lane 3 rDNA.1 PCR negative control, Lane 4 intergenic-region.1, Lane 5 intergenic-region.1 PCR negative control, Lane 6 intergenic-region.2, Lane 7 intergenic-region.1 PCR negative control, Lane 8 intergenic-region.2, Lane 9 intergenic-region.2 PCR negative control



Appendix 5.10. Multiple sequence alignment: 5 tick species *A. americanum, A. glauerti, A. variegatum, A. tuberculatum, A. maculatum* and *R. microplus*; 2 fly species *D. simulans, D. melanogaster*; tick host *B. taurus* and *E. coli* 16S. In *E. coli* 16S protein binding sites are highlighted S7_S9_S19 complex (red), S8_S15_S17 complex (green), S8_S17 complex (aqua) (Weiner et al. 1988)

References

1. Sullivan DE, Gabbard JL, Jr., Shukla M, Sobral B: **Data integration for dynamic and sustainable systems biology resources: challenges and lessons learned**. *Chem Biodivers* 2010, **7**(5):1124-1141.
2. **A Short History of Bioinformatics** [http://www.netsci.org/Science/Bioinform/feature06.html]
3. Rusk N: **Focus on next-generation sequencing data analysis. Forward**. *Nat Methods* 2009, **6**(11 Suppl):S1.
4. Flicek P, Birney E: **Sense from sequence reads: methods for alignment and assembly**. *Nature Methods Supplement* 2009, **6**(11):6-12.
5. Smedley D, Swertz MA, Wolstencroft K, Proctor G, Zouberakis M, Bard J, Hancock JM, Schofield P: **Solutions for data integration in functional genomics: a critical assessment and case study**. *Brief Bioinform* 2008, **9**(6):532-544.
6. Romano P: **Automation of in-silico data analysis processes through workflow management systems**. *Brief Bioinform* 2008, **9**(1):57-68.
7. Wolfsberg TG, Wetterstrand KA, Guyer MS, Collins FS, Baxevanis AD: **A user's guide to the human genome**. *Nat Genet* 2002, **32 Suppl**:1-79.
8. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler DB: **The human genome browser at UCSC**. *Genome Res* 2002, **12**(6):996-1006.
9. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2008, **36**(Database issue):D13-21.
10. Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV: **The Ensembl Web site: mechanics of a genome browser**. *Genome Res* 2004, **14**(5):951-955.
11. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S *et al*: **Ensembl's 10th year**. *Nucleic Acids Res* 2009, **38**(Database issue):D557-562.
12. McPherson JD: **Next-generation gap**. *Nat Methods* 2009, **6**(11 Suppl):S2-5.
13. Lee GW, Kim S: **Genome data mining for everyone**. *BMB Rep* 2008, **41**(11):757-764.
14. Wilkinson MD, Senger M, Kawas E, Bruskiewich R, Gouzy J, Noirot C, Bardou P, Ng A, Haase D, Saiz Ede A *et al*: **Interoperability with Moby 1.0--it's better than sharing your toothbrush!** *Brief Bioinform* 2008, **9**(3):220-231.
15. Goble C, Stevens R: **State of the nation in data integration for bioinformatics**. *J Biomed Inform* 2008, **41**(5):687-693.
16. DiBernardo M, Pottinger R, Wilkinson M: **Semi-automatic web service composition for the life sciences using the BioMoby semantic web framework**. *J Biomed Inform* 2008, **41**(5):837-847.

17. Kawas E, Senger M, Wilkinson MD: **BioMoby extensions to the Taverna workflow management and enactment software**. *BMC Bioinformatics* 2006, **7**:523.
18. Wilkinson M: **Gbrowse Moby: a Web-based browser for BioMoby Services**. *Source Code Biol Med* 2006, **1**:4.
19. Good BM, Wilkinson MD: **The Life Sciences Semantic Web is full of creeps!** *Brief Bioinform* 2006, **7**(3):275-286.
20. Wilkinson MD, Links M: **BioMOBY: an open source biological web services proposal**. *Brief Bioinform* 2002, **3**(4):331-341.
21. Schoof H, Ernst R, Mayer KF: **The PlaNet Consortium: A Network of European Plant Databases Connecting Plant Genome Data in an Integrated Biological Knowledge Resource**. *Comp Funct Genomics* 2004, **5**(2):184-189.
22. **The Gene Ontology in 2010: extensions and refinements**. *Nucleic Acids Res*, **38**(Database issue):D331-335.
23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**(1):25-29.
24. Carrere S, Gouzy J: **REMORA: a pilot in the ocean of BioMoby web-services**. *Bioinformatics* 2006, **22**(7):900-901.
25. Song YC, Kawas E, Good BM, Wilkinson MD, Tebbutt SJ: **DataBiNS: a BioMoby-based data-mining workflow for biological pathways and non-synonymous SNPs**. *Bioinformatics* 2007, **23**(6):780-782.
26. Neron B, Menager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C: **Mobyle: a new full web bioinformatics framework**. *Bioinformatics* 2009, **25**(22):3005-3011.
27. Gordon PM, Trinh Q, Sensen CW: **Semantic Web Service provision: a realistic framework for Bioinformatics programmers**. *Bioinformatics* 2007, **23**(9):1178-1180.
28. Kerhornou A, Guigo R: **BioMoby web services to support clustering of co-regulated genes based on similarity of promoter configurations**. *Bioinformatics* 2007, **23**(14):1831-1833.
29. Ramirez S, Munoz-Merida A, Karlsson J, Garcia M, Perez-Pulido AJ, Claros MG, Trelles O: **MOWServ: a web client for integration of bioinformatic resources**. *Nucleic Acids Res* 2010, **38 Suppl**:W671-676.
30. Katayama T, Arakawa K, Nakao M, Ono K, Aoki-Kinoshita KF, Yamamoto Y, Yamaguchi A, Kawashima S, Chun HW, Aerts J *et al*: **The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium**. *J Biomed Semantics* 2010, **1**(1):8.
31. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P *et al*: **EMBL Nucleotide Sequence Database in 2006**. *Nucleic Acids Res* 2007, **35**(Database issue):D16-20.
32. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. *Nucleic Acids Research* 2006, **34**(Database Issue):D16–D20.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

34. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**(22):4673-4680.
35. Pearson WR: **Rapid and sensitive sequence comparison with FASTP and FASTA**. *Methods Enzymol* 1990, **183**:63-98.
36. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles--database and tools**. *Nucleic Acids Res* 2005, **33**(Database issue):D562-566.
37. Baxevanis AD: **Searching NCBI databases using Entrez**. *Curr Protoc Bioinformatics* 2008, **Chapter 1**:Unit 1 3.
38. Valentin F, Squizzato S, Goujon M, McWilliam H, Paern J, Lopez R: **Fast and efficient searching of biological data resources--using EB-eye**. *Brief Bioinform* 2010, **11**(4):375-384.
39. Kanehisa M: **The KEGG database**. *Novartis Found Symp* 2002, **247**:91-101; discussion 101-103, 119-128, 244-152.
40. Fujii Y, Imanishi T, Gojobori T: **H-Invitational Database: integrated database of human genes**. *Tanpakushitsu Kakusan Koso* 2004, **49**(11 Suppl):1937-1943.
41. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: **BioMart Central Portal--unified access to biological data**. *Nucleic Acids Res* 2009, **1**(37):23-27.
42. Sherman BT, Huang da W, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis**. *BMC Bioinformatics* 2007, **8**:426.
43. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics,* 2003, **4**(41).
44. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD: **PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium**. *Nucleic Acids Res* 2010, **38**(Database issue):D204-210.
45. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J: **The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species**. *Nucleic Acids Res* 2001, **29**(1):159-164.
46. **The ENCODE (ENCyclopedia Of DNA Elements) Project**. *Science* 2004, **306**(5696):636-640.
47. **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.
48. Lister R, Ecker JR: **Finding the fifth base: genome-wide sequencing of cytosine methylation**. *Genome Res* 2009, **19**(6):959-966.
49. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology**. *Nat Rev Genet* 2009, **10**(10):669-680.
50. Collas P: **The current state of chromatin immunoprecipitation**. *Mol Biotechnol*, **45**(1):87-100.

51. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot**. *Methods Mol Biol* 2007, **406**:89-112.
52. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database**. *Nat Genet* 2004, **36**(5):431-432.
53. Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ: **The HUGO Gene Nomenclature Database, 2006 updates**. *Nucleic Acids Res* 2006, **34**(Database issue):D319-321.
54. Hancock JM, Adams NC, Aidinis V, Blake A, Bogue M, Brown SD, Chesler EJ, Davidson D, Duran C, Eppig JT *et al*: **Mouse Phenotype Database Integration Consortium: integration [corrected] of mouse phenome data resources**. *Mamm Genome* 2007, **18**(3):157-163.
55. Zhang Y, De S, Garner JR, Smith K, Wang SA, Becker KG: **Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information**. *BMC Med Genomics* 2010, **3**:1.
56. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM**. *Am J Hum Genet* 2007, **80**(4):588-604.
57. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res* 2002, **30**(1):52-55.
58. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM)**. *Hum Mutat* 2000, **15**(1):57-61.
59. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Res* 2005, **33**(Database issue):D514-517.
60. Pearson P, Francomano C, Foster P, Bocchini C, Li P, McKusick V: **The status of online Mendelian inheritance in man (OMIM) medio 1994**. *Nucleic Acids Res* 1994, **22**(17):3470-3473.
61. Sadasivam RS, Sundar G, Vaughan LK, Tanik MM, Arnett DK: **Genetic region characterization (Gene RECQuest) - software to assist in identification and selection of candidate genes from genomic regions**. *BMC Res Notes* 2009, **2**:201.
62. Fokkema IF, den Dunnen JT, Taschner PE: **LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach**. *Hum Mutat* 2005, **26**(2):63-68.
63. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A *et al*: **IntAct: an open source molecular interaction database**. *Nucleic Acids Res* 2004, **32**(Database issue):D452-455.
64. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R *et al*: **IntAct--open source resource for molecular interaction data**. *Nucleic Acids Res* 2007, **35**(Database issue):D561-565.
65. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J *et al*: **The IntAct molecular interaction database in 2010**. *Nucleic Acids Res* 2010, **38**(Database issue):D525-531.

66. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B *et al*: **Reactome: a database of reactions, pathways and biological processes**. *Nucleic Acids Res* 2010.
67. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L *et al*: **Reactome: a knowledge base of biologic pathways and processes**. *Genome Biol* 2007, **8**(3):R39.
68. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L *et al*: **Reactome: a knowledgebase of biological pathways**. *Nucleic Acids Res* 2005, **33**(Database issue):D428-432.
69. Beloqui A, Guazzaroni ME, Pazos F, Vieites JM, Godoy M, Golyshina OV, Chernikova TN, Waliczek A, Silva-Rocha R, Al-Ramahi Y *et al*: **Reactome array: forging a link between metabolome and genome**. *Science* 2009, **326**(5950):252-257.
70. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B *et al*: **Reactome knowledgebase of human biological pathways and processes**. *Nucleic Acids Res* 2009, **37**(Database issue):D619-622.
71. Stein LD: **Using the Reactome database**. *Curr Protoc Bioinformatics* 2004, **Chapter 8**:Unit 8 7.
72. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M *et al*: **STRING 8--a global view on proteins and their functional interactions in 630 organisms**. *Nucleic Acids Res* 2009, **37**(Database issue):D412-416.
73. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P *et al*: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored**. *Nucleic Acids Res*.
74. Taboada B, Verde C, Merino E: **High accuracy operon prediction method based on STRING database scores**. *Nucleic Acids Res*, **38**(12):e130.
75. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins**. *Nucleic Acids Res* 2003, **31**(1):258-261.
76. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: **STRING 7--recent developments in the integration and prediction of protein interactions**. *Nucleic Acids Res* 2007, **35**(Database issue):D358-362.
77. Bult CJ, Blake JA, Richardson JE, Kadin JA, Eppig JT, Baldarelli RM, Barsanti K, Baya M, Beal JS, Boddy WJ *et al*: **The Mouse Genome Database (MGD): integrating biology with the genome**. *Nucleic Acids Res* 2004, **32**(Database issue):D476-481.
78. Shaw D: **Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype**. *Curr Protoc Bioinformatics* 2004, **Chapter 1**:Unit 1 7.
79. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W *et al*: **The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information**

retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* 2001, **29**(1):102-105.
80. Garcia-Hernandez M, Berardini TZ, Chen G, Crist D, Doyle A, Huala E, Knee E, Lambrecht M, Miller N, Mueller LA *et al*: **TAIR: a resource for integrated Arabidopsis data**. *Funct Integr Genomics* 2002, **2**(6):239-253.
81. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M *et al*: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community**. *Nucleic Acids Res* 2003, **31**(1):224-228.
82. Weems D, Miller N, Garcia-Hernandez M, Huala E, Rhee SY: **Design, implementation and maintenance of a model organism database for Arabidopsis thaliana**. *Comp Funct Genomics* 2004, **5**(4):362-369.
83. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L *et al*: **The Arabidopsis Information Resource (TAIR): gene structure and function annotation**. *Nucleic Acids Res* 2008, **36**(Database issue):D1009-1014.
84. Poole RL: **The TAIR database**. *Methods Mol Biol* 2007, **406**:179-212.
85. Jaiswal P, Ware D, Ni J, Chang K, Zhao W, Schmidt S, Pan X, Clark K, Teytelman L, Cartinhour S *et al*: **Gramene: development and integration of trait and gene ontologies for rice**. *Comp Funct Genomics* 2002, **3**(2):132-136.
86. Jaiswal P, Ni J, Yap I, Ware D, Spooner W, Youens-Clark K, Ren L, Liang C, Zhao W, Ratnapu K *et al*: **Gramene: a bird's eye view of cereal genomes**. *Nucleic Acids Res* 2006, **34**(Database issue):D717-723.
87. Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A *et al*: **Gramene: a growing plant comparative genomics resource**. *Nucleic Acids Res* 2008, **36**(Database issue):D947-953.
88. Ware D: **Gramene**. *Methods Mol Biol* 2007, **406**:315-329.
89. Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S *et al*: **Gramene: a resource for comparative grass genomics**. *Nucleic Acids Res* 2002, **30**(1):103-105.
90. Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K *et al*: **Gramene, a tool for grass genomics**. *Plant Physiol* 2002, **130**(4):1606-1613.
91. Gelbart WM, Crosby M, Matthews B, Rindone WP, Chillemi J, Russo Twombly S, Emmert D, Ashburner M, Drysdale RA, Whitfield E *et al*: **FlyBase: a Drosophila database. The FlyBase consortium**. *Nucleic Acids Res* 1997, **25**(1):63-66.
92. Drysdale R: **Phenotypic data in FlyBase**. *Brief Bioinform* 2001, **2**(1):68-80.
93. Drysdale RA, Crosby MA: **FlyBase: genes and gene models**. *Nucleic Acids Res* 2005, **33**(Database issue):D390-395.
94. Drysdale R: **FlyBase : a database for the Drosophila research community**. *Methods Mol Biol* 2008, **420**:45-59.
95. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R *et al*: **FlyBase: enhancing Drosophila Gene Ontology annotations**. *Nucleic Acids Res* 2009, **37**(Database issue):D555-559.

96. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E *et al*: **VectorBase: a home for invertebrate vectors of human pathogens**. *Nucleic Acids Res* 2007, **35**(Database issue):D503-505.
97. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E *et al*: **VectorBase: a data resource for invertebrate vector genomics**. *Nucleic Acids Res* 2009, **37**(Database issue):D583-587.
98. Megy K, Hammond M, Lawson D, Bruggner RV, Birney E, Collins FH: **Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase**. *Infect Genet Evol* 2009, **9**(3):308-313.
99. Topalis P, Tzavlaki C, Vestaki K, Dialynas E, Sonenshine DE, Butler R, Bruggner RV, Stinson EO, Collins FH, Louis C: **Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase**. *Insect Mol Biol* 2008, **17**(1):87-89.
100. St Pierre S, McQuilton P: **Inside FlyBase: biocuration as a career**. *Fly (Austin)* 2009, **3**(1):112-114.
101. Hunter A, Macgregor A, Szabo T, Takayama N, Schibeci D, Wellington C, Bellgard M: **Yabi: A sophisticated online research environment for Grid, High Performance and Cloud computing**. In.: Murdoch University; 2011.
102. Cieslik M, Mura C: **A lightweight, flow-based toolkit for parallel and distributed bioinformatics pipelines**. *BMC Bioinformatics*, **12**:61.
103. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biol*, **11**(8):R86.
104. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T: **Taverna: a tool for building and running workflows of services**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W729-732.
105. Chen YP, Chen F: **Identifying targets for drug discovery using bioinformatics**. *Expert Opin Ther Targets* 2008, **12**(4):383-389.
106. Chen YP. P, Chen F: **Identifying targets for drug discovery using bioinformatics**. *Expert Opin Ther Targets* 2008, **12**(4):383-389.
107. Saeys Y, Inza I, Larranaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.
108. Yu B: ***In silico* gene discovery**. *Methods Mol Med* 2008, **141**:1-22.
109. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB: **Annotating non-coding regions of the genome**. *Nat Rev Genet* 2010, **11**(8):559-571.
110. Friedrich T, Pils B, Dandekar T, Schultz J, Muller T: **Modelling interaction sites in protein domains with interaction profile hidden Markov models**. *Bioinformatics* 2006, **22**(23):2851-2857.
111. Eddy SR: **Hidden Markov models**. *Curr Opin Struct Biol* 1996, **6**(3):361-365.
112. **HMMER** [http://hmmer.janelia.org/]
113. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2008, **36**(Database issue):D281-288.
114. Schuster-Bockler B, Bateman A: **An introduction to hidden Markov models**. *Curr Protoc Bioinformatics* 2007, **Appendix 3**:Appendix 3A.

115. Cooper BS: **Confronting models with data**. *J Hosp Infect* 2007, **65 Suppl 2**:88-92.
116. Guan Y, Ackert-Bicknell CL, Kell B, Troyanskaya OG, Hibbs MA: **Functional genomics complements quantitative genetics in identifying disease-gene associations**. *PLoS Comput Biol* 2010, **6**(11):e1000991.
117. Flower DR, Doytchinova IA: **Immunoinformatics and the prediction of immunogenicity**. *Appl Bioinformatics* 2002, **1**(4):167-176.
118. Sollner J, Heinzel A, Summer G, Fechete R, Stipkovits L, Szathmary S, Mayer B: **Concept and application of a computational vaccinology workflow**. *Immunome Res* 2010, **6 Suppl 2**:S7.
119. Huang J, Honda W: **CED: a conformational epitope database**. *BMC Immunol* 2006, **7**:7.
120. Parker JM, Guo D, Hodges RS: **New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites**. *Biochemistry* 1986, **25**(19):5425-5432.
121. Karplus PA, Schulz GE: **Prediction of Chain Flexibility in Proteins - A tool for the Selection of Peptide Antigens**. *Naturwissenschafren* 1985, **72**:212-213.
122. Emini EA, Hughes JV, Perlow DS, Boger J: **Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide**. *J Virol* 1985, **55**(3):836-839.
123. Chou PY, Fasman GD: **Prediction of the secondary structure of proteins from their amino acid sequence**. *Adv Enzymol Relat Areas Mol Biol* 1978, **47**:45-148.
124. Kolaskar AS, Tongaonkar PC: **A semi-empirical method for prediction of antigenic determinants on protein antigens**. *FEBS Lett* 1990, **276**(1-2):172-174.
125. Larsen JE, Lund O, Nielsen M: **Improved method for predicting linear B-cell epitopes**. *Immunome Res* 2006, **2**:2.
126. Salimi N, Fleri W, Peters B, Sette A: **Design and utilization of epitope-based databases and predictive tools**. *Immunogenetics* 2010, **62**(4):185-196.
127. Mora M, Telford JL: **Genome-based approaches to vaccine development**. *J Mol Med* 2010, **88**(2):143-147.
128. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements**. *Cytogentic and Genome Research* 2005, **110**:462-467.
129. **RepeatMasker Open-3.0** [http://www.repeatmasker.org]
130. Jurka J, Klonowski P, Dagman V, Pelton P: **CENSOR - a program for identification and elimination of repetitive elements from DNA sequences.** *Computers and Chemistry* 1996, **20**(1):119-121.
131. Makalowski W: **Genomic scrap yard: how genomes utilize all that junk**. *Gene* 2000, **259**:61-67.
132. Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F, Graur D, Ast G: **Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons.** *Mol Cell* 2004, **14**(2):221-231.
133. Price AL, Eskin E, Pevzner PA: **Whole-genome analysis of Alu repeat elements reveals complex evolutionary history.** *Genome Res* 2004, **14**:2245-2252.
134. Yulug IG, Yulug A, Fisher EMC: **The frequency and position of Alu repeats in cDNAs, as determined by database searching**. *Genomics* 1995, **27**:544-548.

135. **H-InvDB** [http://h-invitational.jp/hinv/ahg-db/index.jsp]
136. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank]
137. **FTP site for downloading data files in H-Invitational Database** [ftp://ftp.ddbj.nig.ac.jp/mirror_database/hinv/]
138. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**(1):25-29.
139. **NCBI FTP** [ftp://ftp.ncbi.nih.gov/]
140. **PostgresSQL** [http://www.postgresql.org]
141. **GENE-FTP** [ftp://ftp.ncbi.nlm.nih.gov/gene/DATA]
142. **GO-Downloads** [http://www.geneontology.org/GO.downloads.shtml#ont]
143. Kent WJ: **BLAT-The BLAST-Like Alignment Tool**. *Genome Res* 2002, **12**(4):656-664.
144. Engels R, Yu T, Burge C, Mesirov JP, DeCaprio D, Galagan JE: **Combo: a whole genome comparative browser**. *Bioinformatics* 2006, **22**(14):1782-1783.
145. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER**. *Nucleic Acids Research* 1999, **27**(23):4636-4641.
146. Bras AM, Chatterjee S, Wren BW, Newell DG, Ketley JM: **A novel *Campylobacter jejuni* two-component regulatory system important for temperature-dependent growth and colonization**. *J bacteriol* 1999, **181**:3298-3302.
147. von Mering C, Jensen Jl, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: **STRING 7—recent developments in the integration and prediction of protein interactions**. *Nucleic Acids Res* 2007, **35**(Database issue):D358-D362.
148. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M *et al*: **CDD: specific functional annotation with the Conserved Domain Database**. *Nucleic Acids Res* 2009, **37**(Database issue):D205-210.
149. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H *et al*: **The Bioperl toolkit: Perl modules for the life sciences**. *Genome Res* 2002, **12**(10):1611-1618.
150. Fouts DE, Mongodin EF, Mandrell RE, Miller WG, Rasko DA, Ravel J, Brinkac LM, DeBoy RT, Parker CT, Daugherty SC: **Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species**. *PLoS Biol* 2005, **3**(1):15.
151. Lew-Tabor AE, Moolhuijzen PM, Vance ME, Kurscheid S, Valle MR, Jarrett S, Minchin CM, Jackson LA, Jonsson NN, Bellgard MI *et al*: **Suppressive subtractive hybridization analysis of Rhipicephalus (Boophilus) microplus larval and adult transcript expression during attachment and feeding**. *Vet Parasitol* 2009, **167**(2-4):304-320.
152. Guerrero FD, Miller RJ, Rousseau ME, Sunkara S, Quackenbush J, Lee Y, Nene V: **BmiGI: a database of cDNAs expressed in Boophilus microplus, the tropical/southern cattle tick**. *Insect Biochem Mol Biol* 2005, **35**(6):585-595.
153. Pagel Van Zee J, Geraci NS, Guerrero FD, Wikel SK, Stuart JJ, Nene VM, Hill CA: **Tick genomics: the Ixodes genome project and beyond**. *Int J Parasitol* 2007, **37**(12):1297-1305.

154. **FlyBase: the Drosophila database. The Flybase Consortium**. *Nucleic Acids Res* 1996, **24**(1):53-56.
155. **FlyBase: a Drosophila database. Flybase Consortium**. *Nucleic Acids Res* 1998, **26**(1):85-88.
156. **The FlyBase database of the Drosophila Genome Projects and community literature. The FlyBase Consortium**. *Nucleic Acids Res* 1999, **27**(1):85-88.
157. **The FlyBase database of the Drosophila genome projects and community literature**. *Nucleic Acids Res* 2002, **30**(1):106-108.
158. **The FlyBase database of the Drosophila genome projects and community literature**. *Nucleic Acids Res* 2003, **31**(1):172-175.
159. Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM: **FlyBase: genomes by the dozen**. *Nucleic Acids Res* 2007, **35**(Database issue):D486-491.
160. Grumbling G, Strelets V: **FlyBase: anatomical data, images and queries**. *Nucleic Acids Res* 2006, **34**(Database issue):D484-488.
161. Karamanis N, Seal R, Lewin I, McQuilton P, Vlachos A, Gasperin C, Drysdale R, Briscoe T: **Natural language processing in aid of FlyBase curators**. *BMC Bioinformatics* 2008, **9**:193.
162. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Res* 1998, **8**(3):175-185.
163. Huang X, Madan A: **CAP3: A DNA sequence assembly program**. *Genome Res* 1999, **9**(9):868-877.
164. Mullikin JC, Ning Z: **The phusion assembler**. *Genome Res* 2003, **13**(1):81-90.
165. Chevreux B., Thomas Pfisterer T., Drescher B., Driesel A.J., Müller W.E.G., Wetter T., Suhai S.: **Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs**. *Genome Research* 2004, **14**(6):1147–1159.
166. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.
167. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE *et al*: ***De novo* transcriptome assembly with ABySS**. *Bioinformatics* 2009, **25**(21):2872-2877.
168. Zerbino DR, Birney E: **Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**(5):821-829.
169. Chaisson MJ, Pevzner PA: **Short read fragment assembly of bacterial genomes**. *Genome Res* 2008, **18**(2):324-330.
170. Jurka J, Klonowski P, Dagman V, Pelton P: **CENSOR - a program for identification and elimination of repetitive elements from DNA sequences.** *Computers and Chemistry* 1996, **20**(1):119-121.
171. **RepeatMasker Open-3.0.** [http://www.repeatmasker.org/RMDownload.html]
172. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes**. *Bioinformatics* 2005, **21 Suppl 1**:i351-358.
173. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders**. *Bioinformatics* 2004, **20**(16):2878-2879.
174. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**(1):78-94.

175. Saha S, Raghava GP: **Prediction of continuous B-cell epitopes in an antigen using recurrent neural network**. *Proteins* 2006, **65**(1):40-48.
176. Kulkarni-Kale U, Bhosle S, Kolaskar AS: **CEP: a conformational epitope prediction server**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W168-171.
177. Moolhuijzen P, Kulski JK, Dunn DS, Schibeci D, Barrero R, Gojobori T, Bellgard M: **The transcript repeat element: the human Alu sequence as a component of gene networks influencing cancer**. *Funct Integr Genomics* 2010, **10**(3):307-319.
178. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860-921.
179. Schulz WA, Steinhoff C, Florl AR: **Methylation of endogenous human retroelements in health and disease.** *Curr Top Microbiol Immunol* 2006, **3**(310):211-250.
180. Feschotte C: **Transposable elements and the evolution of regulatory networks**. *Nat Rev Genet* 2008, **9**(5):397-405.
181. Jamalkandi SA, Masoudi-Nejad A: **Reconstruction of Arabidopsis thaliana fully integrated small RNA pathway**. *Funct Integr Genomics* 2009, **9**(4):419-432.
182. Kim VN: **Small RNAs: classification, biogenesis, and function**. *Mol Cells* 2005, **19**(1):1-15.
183. Lev-Maor G, Ram O, Kim E, Sela N, Goren A, Levanon EY, Ast G: **Intronic Alus influence alternative splicing**. *PLoS Genet* 2008, **4**(9):e1000204.
184. Britten RJ, Davidson EH: **Gene regulation for higher cells: a theory**. *Science* 1969, **165**(891):349-357.
185. Britten RJ, Davidson EH: **Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty**. *Q Rev Biol* 1971, **46**(2):111-138.
186. Johnson R, Gamblin RJ, Ooi L, Bruce AW, Donaldson IJ, Westhead DR, Wood IC, Jackson RM, Buckley NJ: **Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication**. *Nucleic Acids Res* 2006, **34**(14):3862-3877.
187. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB: **Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos**. *Dev Cell* 2004, **7**(4):597-606.
188. Novikova O: **Chromodomains and LTR retrotransposons in plants**. *Commun Integr Biol* 2009, **2**(2):158-162.
189. Mattick JS: **A new paradigm for developmental biology**. *J Exp Biol* 2007, **210**(Pt 9):1526-1547.
190. Aravin AA, Hannon GJ, Brennecke J: **The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race**. *Science* 2007, **318**(5851):761-764.
191. Slotkin RK, Martienssen R: **Transposable elements and the epigenetic regulation of the genome**. *Nat Rev Genet* 2007, **8**(4):272-285.
192. Smalheiser NR, Torvik VI: **Mammalian microRNAs derived from genomic repeats**. *Trends Genet* 2005, **21**(6):322-326.
193. Piriyapongsa J, Marino-Ramirez L, Jordan IK: **Origin and evolution of human microRNAs from transposable elements**. *Genetics* 2007, **176**(2):1323-1337.

194. Smalheiser NR, Torvik VI: **Alu elements within human mRNAs are probable microRNA targets**. *Trends Genet* 2006, **22**(10):532-536.
195. Morris KV, Chan SW, Jacobsen SE, Looney DJ: **Small interfering RNA-induced transcriptional gene silencing in human cells**. *Science* 2004, **305**(5688):1289-1292.
196. Chen K, Rajewsky N: **The evolution of gene regulation by transcription factors and microRNAs**. *Nat Rev Genet* 2007, **8**(2):93-103.
197. Grewal SI, Jia S: **Heterochromatin revisited**. *Nat Rev Genet* 2007, **8**(1):35-46.
198. Chandler VL, Stam M: **Chromatin conversations: mechanisms and implications of paramutation**. *Nat Rev Genet* 2004, **5**(7):532-544.
199. Chandler V, Alleman M: **Paramutation: epigenetic instructions passed across generations**. *Genetics* 2008, **178**(4):1839-1844.
200. Stam M, Belele C, Dorweiler JE, Chandler VL: **Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation**. *Genes Dev* 2002, **16**(15):1906-1918.
201. Kawasaki H, Taira K, Morris KV: **siRNA induced transcriptional gene silencing in mammalian cells**. *Cell Cycle* 2005, **4**(3):442-448.
202. Meylan S, Trono D: **Innate immunity against retroviral pathogens: from an ambiguous genetic self to novel therapeutic approaches**. *Swiss Med Wkly* 2009, **139**(49-50):706-711.
203. Goodier JL, Kazazian HH, Jr.: **Retrotransposons revisited: the restraint and rehabilitation of parasites**. *Cell* 2008, **135**(1):23-35.
204. Hauptmann S, Schmitt WD: **Transposable elements--is there a link between evolution and cancer?** *Med Hypotheses* 2006, **66**(3):580-591.
205. Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y: **Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer**. *Cancer Res* 1992, **52**(3):643-645.
206. Houck CM, Rinehart FP, Schmid CW: **A ubiquitous family of repeated DNA sequences in the human genome.** *J Mol Biol* 1979, **132**(3):289-306.
207. Ullu E, Tschudi C: **Alu sequences are processed 7SL RNA genes.** *Nature* 1984, **312**(5990):171-172.
208. Quentin Y: **Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements.** *Nucleic Acids Res* 1992, **20**(13):3397–3401.
209. Quentin Y: **Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes**. *Nucleic Acids Res* 1992, **20**(3):487–493.
210. Kapitonov V, Jurka J: **The age of Alu subfamilies**. *J Mol Evol* 1996, **42**:59-65.
211. Chu WM, Ballard R, Carpick BW, Williams BRG, Schmid CW: **Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR.** *Mol Cell Biol* 1998, **18**(1):58-68.
212. Hagan CR, Sheffield RF, Rudin CM: **Human Alu element retrotransposition induced by genotoxic stress.** *Nat Genet* 2003, **35**(3):219-220.
213. Hasler J, Strub K: **Alu elements as regulators of gene expression.** *Nucleic Acids Res* 2006, **34**(19):5491-5497.

214. Vila MR, Gelpi C, Nicolas A, Morote J, Schwartz SJ, Schwartz S, Meseguer A: **Higher processing rates of Alu-containing sequences in kidney tumours and cell lines with overexpressed Alu-mRNAs.** *Oncol Rep* 2003, **10**(6):1903-1909.
215. Sorek R, Ast G, Graur D: **Alu-containing exons are alternatively spliced**. *Genome Res* 2002, **12**(7):1060-1067.
216. An HJ, Lee D, Lee KH, Bhak J: **The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3' untranslated regions.** *BMC Genomics* 2004, **5**:97.
217. Krull M, Brosius J, Schmitz J: **Alu-SINE exonization: en route to protein-coding function.** *Mol Biol Evol* 2005, **22**(8):1702-1711.
218. Liu WM, Maraia RJ, Rubin CM, Schmid CW: **Alu transcripts: cytoplasmic localisation and regulation by DNA methylation.** *Nucleic Acids Res* 1994, **22**:1087–1095.
219. Kondo Y, Issa JP: **Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells**. *J Biol Chem* 2003, **278**(30):27658-27662.
220. Saito Y, Suzuki H, Tsugawa H, Nakagawa I, Matsuzaki J, Kanai Y, Hibi T: **Chromatin remodeling at Alu repeats by epigenetic treatment activates silenced microRNA-512-5p with downregulation of Mcl-1 in human gastric cancer cells**. *Oncogene* 2009, **28**(30):2738-2744.
221. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H *et al*: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12**(10):1611-1618.
222. **Perl** [http://www.perl.com]
223. **GMOD** [http://www.gmod.org/cmap/index.shtml]
224. Yamasaki C, Koyanagi KO, Fujii Y, Itoh T, Barrero R, Tamura T, Yamaguchi-Kabata Y, Tanino M, Takeda J, Fukuchi S *et al*: **Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB)**. *Gene* 2005, **364**(30):99-107.
225. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M *et al*: **Integrative annotation of 21,037 human genes validated by full-length cDNA clones**. *PLoS Biol* 2004, **2**(6):856-875.
226. Maas S, Patt S, Schrey M, Rich A: **Underediting of glutamine receptor GluR-B mRNA in malignant gliomas.** *Proc Natl Acad Sci* 2001, **98**(25):14687-14692.
227. Paz N, Levanon EY, Amariglio N, Heimberger AB, Ram Z, Constantini S, Barbash ZS, Adamsky K, Safran M, Hirschberg A *et al*: **Altered adenosine-to-inosine RNA editing in human cancer**. *Genome Res* 2007, **17**:1586-1595.
228. Zilberman DE, Safran M, Paz N, Amariglio N, Simon A, Fridman E, Kleinmann N, Ramon J, Rechavi G: **Does RNA editing play a role in the development of urinary bladder cancer?** *Urol Oncol* 2009.
229. **GO-Slims** [http://www.geneontology.org/GO_slims/]
230. Kim DS, Huh JW, Kim HS: **Transposable elements in human cancers by genome-wide EST alignment** *Gene and Genetic Systems* 2007, **82**(2):145-156.
231. Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallengger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D *et al*: **Systematic identification**

of abundant A-to-I editing sites in the human transcriptome. *Nat Biotech* 2004, **22**(8):1001-1005.
232. Kim DDY, Kim TTY, Walsh T, Kobayashi Y, Matise TC, Buyske S, Gabriel A: **Widespread RNA Editing of Embedded Alu Elements in the Human Transcriptome**. *Genome Res* 2004, **14**:1719-1725.
233. Sobczak K, Krzyzosiak WJ: **Structural Determinants of BRCA1 Translational Regulation**. *J Biol Chem* 2002, **277**(19):17349 –17358.
234. Lei H, Vorechovsky I: **Identification of Splicing Silencers and Enhancers in Sense Alus: a Role for Pseudoacceptors in Splice Site Repression**. *Mol Cell Biol* 2005:6912–6920.
235. Lei H, Day INM, Vorechovsky I: **Exonization of AluYa5 in the human ACE gene requires mutations in both 3' and 5' splice sites and is facilitated by a conserved splicing enhancer**. *Nucleic Acids Res* 2005, **33**(12):3897–3906.
236. Claverie-Martin F, Flores C, Anton-Gamero M, Gonzalez-Acosta H, Garcıa-Nieto V: **The Alu insertion in the CLCN5 gene of a patient with Dent's disease leads to exon 11 skipping**. *J Hum Genet* 2005, **50**:370–374.
237. Makalowski W: **Not junk after all.** *Science* 2003, **300**:1246-1247.
238. Britten RJ: **Coding sequences of functioning human genes derived entirely from mobile element sequences**. *Proc Natl Acad Sci* 2004, **101**(48):16825–16830.
239. De La Monte SM, Ghanbari K, Frey WH, Beheshti I, Averback P, Hauser SL, Ghanbari HA, Wands JR: **Characterization of the AD7c-NTP cDNA expression in Alzheimer' disease and measurement of a 41-kD Protein in cerebrospinal fluid**. *J Clin Invest* 1997, **100**:3093– 3104.
240. Kriegs JA, Schmitz J, Makalowski W, Brosius J: **Does the AD7c-NTP locus encode a protein?** . *BBA - Gene Structure and Expression* 2005, **1727**:1-4.
241. Gotea V, Makalowski W: **Do transposable elements really contribute to protease?** *Trends In Genetics* 2006, **22**(5):260-267.
242. Yura K, Shionyu M, Hagino K, Hijikata A, Hirashima Y, Nakahara T, Eguchi T, Shinoda K, Yamaguchi A, Takahashi K *et al*: **Alternative splicing in human transcriptome: functional and structural influence on proteins.** *Gene* 2006, **380**(2):63-71.
243. Yoder JA, Walsh CP, Bestor T.H.: **Cytosine methylation and the ecology of intragenomic parasites**. *Trends Genet* 1997, **13**:335–340.
244. Liu WM, Maraia RJ, Rubin CM, Schmid CW: **Alu transcripts: cytoplasmic localisation and regulation by DNA methylation.** *Nucleic Acids Res* 1994, **22**:1087–1095.
245. Bird AP: **Functions for DNA methylation in vertebrates.** *Cold Spring Harb Symp Quant Biol* 1993, **58**:281–285. .
246. Richards K, Zhang B, Baggerly K, Colella S, Lang J, Schuller D, Krahe R: **Genome-wide hypomethylation in head and neck cancer is more pronounced in HPV-negative tumors and is associated with genomic instability.** *PLoS ONE* 2009, **4**(3):e4941.
247. Bollati V, Fabris S, Pegoraro V, Ronchetti D, Mosca L, Deliliers GL, Motta V, Bertazzi PA, Baccarelli A, Neri A: **Differential repetitive DNA methylation in multiple myeloma molecular subgroups**. *Carcinogenesis* 2009, **30**(8):1330-1335.

248. Daskalos A, Nikolaidis G, Xinarianos G, Savvari P, Cassidy A, Zakopoulou R, Kotsinas A, Gorgoulis V, Field JK, Liloglou T: **Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer.** *Int J Cancer* 2009, **124**(1):81-87.
249. Wilson AS, Power BE, Molloy P: **DNA hypomethylation and human diseases.** *Biochim Biophys Acta* 2007, **1775**(1):138-162.
250. Borchert GM, Lanier W, Davidson B: **RNA polymerase III transcribes human microRNAs**. *Nat Struct Mol Biol* 2006, **13**(12):1097-1101.
251. Lehnert S, Van Loo P, Thilakarathne PJ, Marynen P, Verbeke G, Schuit FC: **Evidence for co-evolution between human microRNAs and Alu-repeats**. *PLoS ONE* 2009, **4**(2):e4456.
252. Moolhuijzen PM, Lew-Tabor AE, Wlodek BM, Aguero FG, Comerci DJ, Ugalde RA, Sanchez DO, Appels R, Bellgard M: **Genomic analysis of *Campylobacter fetus* subspecies: identification of candidate virulence determinants and diagnostic assay targets**. *BMC Microbiol* 2009, **9**:86.
253. Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, La T, Hampson DJ, Bellgard M, Wassenaar TM, Ussery DW: **Ten years of bacterial genome sequencing: comparative-genomics-based discoveries**. *Funct Integr Genomics* 2006, **6**(3):165-185.
254. Dobrindt U, Hochhut B, Hentschel U, Hacker J: **Genomic islands in pathogenic and environmental microorganisms**. *Nat Rev Microbiol* 2004, **2**(5):414-424.
255. Fluit AC, Schmitz FJ, Verhoef J, Milatovic D: **Daptomycin in vitro susceptibility in European Gram-positive clinical isolates**. *Int J Antimicrob Agents* 2004, **24**(1):59-66.
256. Holmes AJ, Holley MP, Mahon A, Nield B, Gillings M, Stokes HW: **Recombination activity of a distinctive integron-gene cassette system associated with Pseudomonas stutzeri populations in soil**. *J Bacteriol* 2003, **185**(3):918-928.
257. Peters ED, Leverstein-van Hall MA, Box AT, Verhoef J, Fluit AC: **Novel gene cassettes and integrons**. *Antimicrob Agents Chemother* 2001, **45**(10):2961-2964.
258. van Belkum A, Scherer S, van Alphen L, Verbrugh H: **Short-sequence DNA repeats in prokaryotic genomes**. *Microbiol Mol Biol Rev* 1998, **62**(2):275-293.
259. Rocha EP, Danchin A, Viari A: **Functional and evolutionary roles of long repeats in prokaryotes**. *Res Microbiol* 1999, **150**(9-10):725-733.
260. Ou HY, Chen LL, Lonnen J, Chaudhuri RR, Thani AB, Smith R, Garton NJ, Hinton J, Pallen M, Barer MR *et al*: **A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria**. *Nucleic Acids Res* 2006, **34**(1):e3.
261. Mahillon J, Leonard C, Chandler M: **IS elements as constituents of bacterial genomes**. *Res Microbiol* 1999, **150**(9-10):675-687.
262. Yang J, Chen L, Sun L, Yu J, Jin Q: **VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics**. *Nucleic Acids Res* 2008, **36**(Database issue):D539-542.
263. Weiss RA: **Virulence and pathogenesis**. *Trends Microbiol* 2002, **10**(7):314-317.
264. Brogden KA, Roth JA, Stanton TB, Bolin CA, Minion FC, Wannemuehler MJ: **Virulence Mechanisms of Bacterial Pathogens**, 3 edn. Washington DC: ASM Press; (2000)

265. Jores J, Rumer L, Wieler LH: **Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli***. *Int J Med Microbiol* 2004, **294**(2-3):103-113.
266. Zagursky RJ, Olmsted SB, Russell DP, Wooters JL: **Bioinformatics: how it is being used to identify bacterial vaccine candidates**. *Expert Rev Vaccines* 2003, **2**(3):417-436.
267. Zagursky RJ, Russell D: **Bioinformatics: use in bacterial vaccine discovery**. *Biotechniques* 2001, **31**(3):636, 638, 640, passim.
268. Serruto D, Adu-Bobie J, Capecchi B, Rappuoli R, Pizza M, Masignani V: **Biotechnology and vaccines: application of functional genomics to Neisseria meningitidis and other bacterial pathogens**. *J Biotechnol* 2004, **113**(1-3):15-32.
269. Serruto D, Galeotti CL: **The signal peptide sequence of a lytic transglycosylase of Neisseria meningitidis is involved in regulation of gene expression**. *Microbiology* 2004, **150**(Pt 5):1427-1437.
270. Pizza M, Giuliani MM, Fontana MR, Douce G, Dougan G, Rappuoli R: **LTK63 and LTR72, two mucosal adjuvants ready for clinical trials**. *Int J Med Microbiol* 2000, **290**(4-5):455-461.
271. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B *et al*: **Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing**. *Science* 2000, **287**(5459):1816-1820.
272. Garcia MM, Eaglesome MD, Rigby C: **Campylobacters important in veterinary medicine**. *Vet Bull* 1983, **53**:793–818.
273. Mshelia GD, Singh J, Amin J. D, Woldehiwet Z, Egwu GO, Murray RD: **Bovine venereal campylobacteriosis: an overview**. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources* 2007, **2**(80):14.
274. McMillen L, Fordyce G, Doogan VJ, Lew AE: **Comparison of Culture and a Novel 5' *Taq* Nuclease Assay for Direct Detection of *Campylobacter fetus* subsp. *venerealis* in Clinical Specimens from Cattle.** *J Clin Microbiol* 2006, **44**:938-945.
275. Parkhill J: **The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences.** *Nature* 2000 **403**(6770):665-668.
276. On SL, Harrington CS: **Evaluation of numerical analysis of PFGE-DNA profiles for differentiating *Campylobacter fetus* subspecies by comparison with phenotypic, PCR and 16S rDNA sequencing methods**. *J Appl Microbiol* 2001, **90**(2):285-293.
277. Leece JG: **Some biochemical characteristics of *Vibrio fetus* and other related Vibrios isolated from animals.** *J Bacteriol* 1958, **76**:312-316.
278. Clark BL, Dufty JH, Monsbourgh MJ: **A method for maintaining the viability of *Vibrio fetus* var. *venerealis* in samples of preputial secretions collected from carrier bulls.** *Aust Vet J* 1972, **48**(8):462-464.
279. Clark BL, Dufty JH: **Isolation of *Campylobacter fetus* from bulls**. *Aust Vet J* 1978, **54**:262-263.
280. Jones RL, Davis MA, Vonbyern H: **Cultural procedures for the isolation of *Campylobacter fetus* subsp. *venerealis* from preputial secretions and the occurrence of antimicrobial resistance**. *Proceedings of the Annual Meeting of the American Association of Veterinary Laboratory Diagnosticians* 1985, **28**:225-238.

281. van Bergen MA, Simons G, van der Graaf-van Bloois L, van Putten JP, Rombout J, Wesley I, Wagenaar JA: **Amplified fragment length polymorphism based identification of genetic markers and novel PCR assay for differentiation of *Campylobacter fetus* subspecies** *J Med Microbiol* 2005, **54**:1217-1224.
282. Chang W, Ogg JE: **Transduction in *Vibrio fetus***. *Am J Vet Res* 1970, **31**:919-924.
283. Chang W, Ogg JE: **Transduction and mutation to glycine tolerance in *Vibrio fetus***. *Am J Vet Res* 1971, **32**:649-653.
284. Veron M, Chatelain R: **Taxonomic Study of the genus *Campylobacter* Sebald and Veron and designation of the neotype strain for the type species. *Campylobacter fetus* (Smith and Taylor) Sebald and Veron**. *Int J Sys Bacteriol* 1973, **23**:122-134.
285. van Bergen MA, Dingle KE, Maiden MC, Newell DG, van der Graaf-Van Bloois L, van Putten JP, Wagenaar JA: **Clonal nature of *Campylobacter fetus* as defined by multilocus sequence typing**. *J Clin Microbiol* 2005, **43**:5888-5898.
286. Schulze F, Bagon A, Muller W, Hotzel H: **Identification of *Campylobacter fetus* subspecies by phenotypic differentiation and PCR**. *J Clin Microbiol* 2006, **44**(6):2019-2024.
287. Hum S, Quinn K, Brunner J, On SL: **Evaluation of a PCR assay for identification and differentiation of *Campylobacter fetus* subspecies**. *Aust Vet J* 1997, **75**:827-831.
288. Abril C, Vilei EM, Brodard I, Burnens A, Frey J, Miserez R: **Discovery of insertion element IS*Cfe*1: a new tool for *Campylobacter fetus* subspecies differentiation.** *Clin Microbiol Infect* 2007, **13**(10):993-1000.
289. Willoughby K, Nettleton PF, Quirie M, Maley M.A, Foster G, Toszeghy M, Newell DG: **A multiplex polymerase chain reaction to detect and differentiate *Campylobacter fetus* subspecies *fetus* and *Campylobacter fetus* -species *venerealis*: use on UK isolates of *C. fetus* and other *Campylobacter* spp.** *J Appl Microbiol* 2005, **99**(4):758-766.
290. Sambrook J, Fritsch EF, Maniatis T: **In Molecular cloning: A laboratory manual**. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.
291. Clark BL, Dufty JH, Monsbourgh MJ, Parsonson IM: **Immunisation against bovine vibriosis due to Campylobacter fetus subsp. fetus biotype intermedius**. *Aust Vet J* 1976, **52**:,362-365.
292. Agüero F, Verdún RE, Frasch AC, Sánchez DO: **A random sequencing approach for the analysis of the *Trypanosoma cruzi* genome: general structure, large gene and repetitive DNA families, and gene discovery**. *Genome Res* 2000, **10**(12):1996-2005.
293. **RefSeq-Genome** [http://www.ncbi.nlm.nih.gov/RefSeq/]
294. Binnewies TT, Hallin PF, Staerfeldt HH, Ussery DW: **Genome Update: proteome comparisons**. *Microbiology* 2005, **151**(Pt 1):1-4.
295. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.
296. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers**. *Methods Mol Biol* 2000, **132**:365-386.

297. Kienesberger S, Gorkiewicz G, Joainig MM, Scheicher SR, Leitner E, Zechner EL: **Development of Experimental Genetic Tools for *Campylobacter fetus*** *Appl Environ Microbiol* 2007, **73**(14):4619-4630.
298. Asakura M, Samosornsuk W, M T, Kobayashi K, Misawa N, Kusumoto M, Nishimura K, Matsuhisa A, Yamasaki S: **Comparative analysis of cytolethal distending toxin (cdt) genes among *Campylobacter jejuni, C. coli* and *C. fetus* strains.** *Microb Pathog* 2007, **42**(5-6):174-183.
299. Lew AE, Guo S-Y, Venus B, Moolhuijzen P, Sanchez D, Trott D, Burrell P, Wlodek B, Bellgard M: **Comparative genome analysis applied to develop novel PCR assays to characterise and identify *Campylobacter fetus* subsp. *venerealis* isolates**. *Zoonoses and Public Health* 2007 **54**(Supplement 1):154.
300. **ATCC** [http://www.atcc.org/]
301. Salama SM, Garcia MM, Taylor DE: **Differentiation of the subspecies of *Campylobacter fetus* by genomic sizing**. *Int J Sys Bacteriol* 1992, **42**: 446-450.
302. Hiett KL, Stintzi A, Andacht TM, Kuntz RL, Seal BS: **Genomic differences between *Campylobacter jejuni* isolates identify surface membrane and flagellar function gene products potentially important for colonizing the chicken intestine**. *Funct Integr Genomics* 2008, **8**:407–420.
303. Vivona S, Gardy Jl, Ramachandran S, Brinkman F, Raghava GPS, Flwer DR, Filippini F: **Computer-aided biotechnology form immuno-informatics to reverse vaccinology**. *Trends in biotechnol* 2007, **26**(4):190-200.
304. Wassenaar TM, Bleumink-Pluym NM, van der Zeijst BA: **Inactivation of *Campylobacter jejuni* flagellin genes by homologous recombination demonstrates that *flaA* but not *flaB* is required for invasion**. *EMBO J* 1991, **10**:2055-2061.
305. Carrillo CD, Taboada E, Nash JHE, Lanthier P, Kelly J, Lau PC, Verhulp R, Mykytczuk O, Sy J, Findlay WA: **Genome-wide expression analyses of *Campylobacter jejuni* NCTC11168 reveals coordinate regulation of motility and virulence by *flhA***. *J Biol Chem* 2004, **279**(19):20327-20338.
306. Yao R, Burr DH, Doig P, Trust TJ, Niu H: **Isolation of motile and non-motile insertional elements of *Campylobacter jejuni*: The role of motility in adherence and invasion of eukaryotic cells**. *Mol Microbiol* 1994, **14**:883-893.
307. Fernando U, Biswas D, Allan B, Willson P, Potter AA: **Influence of *Campylobacter jejuni fliA*, *rpoN* and *flgK* genes on colonization of the chicken gut**. *Int J Food Microbiol* 2007, **118**:194-200.
308. Konkel ME, Klena JD, River-Amill V, Monteville MR, Biswas D, Raphael B, Mickelson J: **Secretion of virulence proteins from *Campylobacter jejuni* is dependent on a functional flagellar export apparatus**. *J Bacteriol* 2004, **186**:3296-3303.
309. Jagannathan A, Constantinidou C, Penn CW: **Roles of *rpoN*, *fliA* and *flgR* in expression of flagella in *Campylobacter jejuni***. *J Bacteriol* 2001, **183**:2937-2942.
310. Mellmann A, Mosters J, Bartelt E, Roggentin P, Ammon A, Friedrich AW, Karch H, Harmsen D: **Sequence-based typing of *flaB* is a more stable screening tool that typing of *flaA* for monitoring of *Campylobacter* populations**. *J Clin Microbiol* 2004, **42**:4840-4842.

311. Konkel ME, Garvis SG, Tipton SL, Anderson DE, Cieplak W: **Identification and molecular cloning of a gene encoding a fibronectin-binding protein (*CadF*) from *Campylobacter jejuni***. *Mol Microbiol* 1997, **24**:953-963.
312. Pei Z, Burucoa C, Grignon B, Baqar S, Huang X-Z, Kopecko DJ, Bourgeois AL, Fauchere J-L, Blaser MJ: **Mutation in the *peb1A* locus of *Campylobacter jejuni* reduces interactions with epithelial cells and intestinal colonization of mice.** *Infect Immun* 1998, **66**:938-943.
313. Jin S, Joe A, Lynett J, Hani EK, Sherman P, Chan VL: ***JlpA*, a novel surface-exposed lipoprotein specific to *Campylobacter jejuni*, mediates adherence to host epithelial cells**. *Mol Microbiol* 2001, **39**:1225-1236.
314. Moser I, Schroeder W, Salnikow J: ***Campylobacter jejuni* major outer membrane protein and a 59-kDa protein are involved in binding to fibronectin and INT 407 cell membranes**. *FEMS Microbiol Letts* 1997, **157**:233-238.
315. Graham LL, Friel T, Woodman RL: **Fibronectin enhances *Campylobacter fetus* interaction with extracellular matrix components and INT 407 cells**. *Can J Microbiol* 2008, **54**:37-47.
316. Jain K, Prasad KN, Sinha S, Husain N: **Differences in virulence attributes between cytolethal distending toxin positive and negative *Campylobacter jejuni* strains**. *J Med Microbiol* 2008, **57**:267-272.
317. Christie PJ, Atmakuri K, Krishnamoorthy V, Jakubowski S, Cascales E: **Biogenesis, architecture and function of bacterial Type IV secretion systems**. *Annu Rev Microbiol* 2005, **59**:451-485.
318. Ebersbach G, Gerdes K: **Plasmid segregation mechanisms**. *Annu Rev Genet* 2005, **39**:453-479.
319. Moolhuijzen P, Lew-Tabor A, Morgan ATJ, Rodriguez Valle M, Peterson GD, Dowd S. E, Guerrero F, Bellgard M, Appels R: **The complexity of *Rhipicephalus* (*Boophilus*) *microplus* genome characterised through detailed analysis of two BAC clones**. *BMC Research Notes* 2011, **4**(254).
320. Patarroyo JH, Portela RW, De Castro RO, Pimentel JC, Guzman F, Patarroyo ME, Vargas MI, Prates AA, Mendes MA: **Immunization of cattle with synthetic peptides derived from the *Boophilus microplus* gut protein (Bm86)**. *Vet Immunol Immunopathol* 2002, **88**(3-4):163-172.
321. Lees K, Woods DJ, Bowman AS: **Transcriptome analysis of the synganglion from the brown dog tick, *Rhipicephalus sanguineus***. *Insect Mol Biol*, **19**(3):273-282.
322. Nolan J, Wilson JT, Green PE, Bird PE: **Synthetic pyrethroid resistance in field samples in the cattle tick (*Boophilus microplus*)**. *Aust Vet J* 1989, **66**(6):179-182.
323. Patarroyo JH, Costa JO: **Susceptibility of Brazilian samples of *Boophilus microplus* to organophosphorus acaricides**. *Trop Anim Health Prod* 1980, **12**(1):6-10.
324. Ramakrishnan VG, Aljamali MN, Sauer JR, Essenberg RC: **Application of RNA interference in tick salivary gland research**. *J Biomol Tech* 2005, **16**(4):297-305.
325. Waxman L, Smith DE, Arcuri KE, Vlasuk GP: **Tick anticoagulant peptide (TAP) is a novel inhibitor of blood coagulation factor Xa**. *Science* 1990, **248**(4955):593-596.

326. Aljamali M, Bowman AS, Dillwith JW, Tucker JS, Yates GW, Essenberg RC, Sauer JR: **Identity and synthesis of prostaglandins in the lone star tick, Amblyomma americanum (L.), as assessed by radio-immunoassay and gas chromatography/mass spectrometry**. *Insect Biochem Mol Biol* 2002, **32**(3):331-341.
327. Bergman DK, Palmer MJ, Caimano MJ, Radolf JD, Wikel SK: **Isolation and molecular cloning of a secreted immunosuppressant protein from Dermacentor andersoni salivary gland**. *J Parasitol* 2000, **86**(3):516-525.
328. Paesen GC, Adams PL, Harlos K, Nuttall PA, Stuart DI: **Tick histamine-binding proteins: isolation, cloning, and three-dimensional structure**. *Mol Cell* 1999, **3**(5):661-671.
329. Paesen GC, Adams PL, Nuttall PA, Stuart DL: **Tick histamine-binding proteins: lipocalins with a second binding cavity**. *Biochim Biophys Acta* 2000, **1482**(1-2):92-101.
330. Pruett JH: **Immunological control of arthropod ectoparasites--a review**. *Int J Parasitol* 1999, **29**(1):25-32.
331. Neuath AR, Kent SBH: **Requirements for successfully synthetic peptide vaccines**. *Ann Inst Pasteur/Virol* 1986, **E 137**:513-514.
332. Sharp PJ, McInerney BV, Smith DR, Turnbull IF, Kemp DH, Rand KN, Cobon GS: **Chromatography and generation of specific antisera to synthetic peptides from a protective *Boophilus microplus* antigen**. *J Chromatogr* 1990, **512**:189-202.
333. Willadsen P: **Immunological control of ectoparasites: past achievements and future research priorities**. *Genet Anal* 1999, **15**(3-5):131-137.
334. Trimnell AR, Hails RS, Nuttall PA: **Dual action ectoparasite vaccine targeting 'exposed' and 'concealed' antigens**. *Vaccine* 2002, **20**(29-30):3560-3568.
335. Willadsen P: **The molecular revolution in the development of vaccines against ectoparasites**. *Vet Parasitol* 2001, **101**(3-4):353-368.
336. Mulenga A, Sugimoto C, Onuma M: **Issues in tick vaccine development: identification and characterization of potential candidate vaccine antigens**. *Microbes Infect* 2000, **2**(11):1353-1361.
337. Rand KN, Moore T, Sriskantha A, Spring K, Tellam R, Willadsen P, Cobon GS: **Cloning and expression of a protective antigen from the cattle tick *Boophilus microplus***. *Proc Natl Acad Sci U S A* 1989, **86**(24):9657-9661.
338. de la Fuente J, Rodriguez M, Redondo M, Montero C, Garcia-Garcia JC, Mendez L, Serrano E, Valdes M, Enriquez A, Canales M *et al*: **Field studies and cost-effectiveness analysis of vaccination with Gavac against the cattle tick *Boophilus microplus***. *Vaccine* 1998, **16**(4):366-373.
339. Saldivar L, Guerrero FD, Miller RJ, Bendele KG, Gondro C, Brayton KA: **Microarray analysis of acaricide-inducible gene expression in the southern cattle tick, Rhipicephalus (Boophilus) microplus**. *Insect Mol Biol* 2008, **17**(6):597-606.
340. Reck J, Berger M, Marks FS, Zingali RB, Canal CW, Ferreira CA, Guimaraes JA, Termignoni C: **Pharmacological action of tick saliva upon haemostasis and the neutralization ability of sera from repeatedly infested hosts**. *Parasitology* 2009:1-11.

341. Piper EK, Jonsson NN, Gondro C, Lew-Tabor AE, Moolhuijzen P, Vance ME, Jackson LA: **Immunological profiles of Bos taurus and Bos indicus cattle infested with the cattle tick, Rhipicephalus (Boophilus) microplus**. *Clin Vaccine Immunol* 2009, **16**(7):1074-1086.
342. Castelli E, Caputo V, Morello V, Tomasino RM: **Local reactions to tick bites**. *Am J Dermatopathol* 2008, **30**(3):241-248.
343. Parizi LF, Pohl PC, Masuda A, Vaz Junior Ida S: **New approaches toward anti-Rhipicephalus (Boophilus) microplus tick vaccine**. *Rev Bras Parasitol Vet* 2009, **18**(1):1-7.
344. Harrington D, Canales M, de la Fuente J, de Luna C, Robinson K, Guy J, Sparagano O: **Immunisation with recombinant proteins subolesin and Bm86 for the control of Dermanyssus gallinae in poultry**. *Vaccine* 2009, **27**(30):4056-4063.
345. Kumar A, Garg R, Yadav CL, Vatsya S, Kumar RR, Bedarkar SN: **Immune responses against recombinant tick antigen, Bm95, for the control of Rhipicephalus (Boophilus) microplus ticks in cattle**. *Vet Parasitol* 2009.
346. Kemp DH, Stone BF, Binnington KC: **Tick attachment and feeding: role of the mouthparts, feeding apparatus, salivary gland secretions and host response**. In: *Physiology of ticks.* Edited by Obenchain FD, Galun R: Oxford: Pergamon Press; 1982: 119–168.
347. Brown SJ, Shapiro SZ, Askenase PW: **Characterization of tick antigens inducing host immune resistance. I. Immunization of guinea pigs with Amblyomma americanum-derived salivary gland extracts and identification of an important salivary gland protein antigen with guinea pig anti-tick antibodies**. *J Immunol* 1984, **133**(6):3319-3325.
348. Gillet L, Schroeder H, Mast J, Thirion M, Renauld JC, Dewals B, Vanderplasschen A: **Anchoring tick salivary anti-complement proteins IRAC I and IRAC II to membrane increases their immunogenicity**. *Vet Res* 2009, **40**(5):51.
349. Menten-Dedoyart C, Couvreur B, Thellin O, Drion PV, Herry M, Jolois O, Heinen E: **Influence of the Ixodes ricinus tick blood-feeding on the antigen-specific antibody response in vivo**. *Vaccine* 2008, **26**(52):6956-6964.
350. McCosker PJ: **Global aspects of the management and control of ticks of veterinary importance**. *Recent Adv Acarol* 1979, **2**:45–53.
351. George JE, Davey RB, Pound JM: **Introduced ticks and tick-borne diseases: the threat and approaches to eradication**. *Vet Clin North Am Food Anim Pract* 2002, **18**(3):401-416, vi.
352. Palmer MJ, Bantle JA, Guo X, Fargo WS: **Genome size and organization in the ixodid tick Amblyomma americanum (L.)**. *Insect Mol Biol* 1994, **3**(1):57-62.
353. Ullmann AJ, Lima CM, Guerrero FD, Piesman J, Black WCt: **Genome size and organization in the blacklegged tick, Ixodes scapularis and the Southern cattle tick, Boophilus microplus**. *Insect Mol Biol* 2005, **14**(2):217-222.
354. Geraci NS, Spencer Johnston J, Paul Robinson J, Wikel SK, Hill CA: **Variation in genome size of argasid and ixodid ticks**. *Insect Biochem Mol Biol* 2007, **37**(5):399-408.
355. Sunter JD, Patel SP, Skilton RA, Githaka N, Knowles DP, Scoles GA, Nene V, de Villiers E, Bishop RP: **A novel SINE family occurs frequently in both genomic**

DNA and transcribed sequences in ixodid ticks of the arthropod sub-phylum Chelicerata. *Gene* 2008, **415**(1-2):13-22.
356. Kidwell MG: **Transposable elements and the evolution of genome size in eukaryotes**. *Genetica* 2002, **115**(1):49-63.
357. Okada N: **SINEs**. *Curr Opin Genet Dev* 1991, **1**(4):498-504.
358. Ullu E, Tschudi C: **Alu sequences are processed 7SL RNA genes**. *Nature* 1984, **312**(5990):171-172.
359. Hill CA, Wikel SK: **The Ixodes scapularis Genome Project: an opportunity for advancing tick research**. *Trends Parasitol* 2005, **21**(4):151-153.
360. Nene V: **Tick genomics--coming of age**. *Front Biosci* 2009, **14**:2666-2673.
361. Wang M, Guerrero FD, Pertea G, Nene VM: **Global comparative analysis of ESTs from the southern cattle tick, Rhipicephalus (Boophilus) microplus**. *BMC Genomics* 2007, **8**:368.
362. Guerrero FD, Moolhuijzen PM, Peterson DG, Bidwell S, Caler E, Appels R, Bellgard M, Nene VM, Djikeng A: **Reassociation kinetics-based approach for partial genome sequencing of the cattle tick, Rhipicephalus (Boophilus) microplus**. *BMC Genomics* 2010, **11**:374.
363. **Sequencing of BAC ends from a *Rhipicephalus microplus* BAC library** [http://www.ars.usda.gov/research/projects/projects.htm?ACCN_NO=415116]
364. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**(3):186-194.
365. Bellgard MI, Moolhuijzen PM, Guerrero F.D., Appels R, Schibeci D, Rodriguez-Valle M, Barrero R, Hunter A, Lew-Tabor AE: **CattleTickBase: Internet-based analysis tools and bioinformatics repository of available genomics resources for *Rhipicephalus* (*Boophilus*) *microplus***. *International Journal for Parasitology* 2011, **In review**.
366. Davey RB, Garza Jr. J, Thompson GD, Drummond RO: **Ovipositional biology of the cattle tick, Boophilus annulatus (Acari: Ixodidae), in the laboratory**. *J Med Entomol* 1980(17):287-289.
367. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.
368. Sonnhammer ELL, Durbin R: **A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis**. *Gene* 1995, **167**:1-10
369. Ribeiro JM, Alarcon-Chaidez F, Francischetti IM, Mans BJ, Mather TN, Valenzuela JG, Wikel SK: **An annotated catalog of salivary gland transcripts from Ixodes scapularis ticks**. *Insect Biochem Mol Biol* 2006, **36**(2):111-129.
370. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank]
371. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H *et al*: **The Bioperl toolkit: Perl modules for the life sciences**. *Genome Res* 2002, **12**(10):1611-1618.
372. Hunter A, Schibeci D, Hiew HL, Bellgard M: **Grendel: A bioinformatics Web Service-based architecture for accessing HPC resources**. *Australasian Workshop on Grid Computing and e-Research* 2005.
373. Koski LB, Gray MW, Lang BF, Burger G: **AutoFACT: an automatic functional annotation and classification tool**. *BMC Bioinformatics* 2005, **6**:151.

374. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ: **Jalview Version 2--a multiple sequence alignment editor and analysis workbench**. *Bioinformatics* 2009, **25**(9):1189-1191.
375. **PHYLIP (the PHYLogeny Inference Package)** [http://evolution.genetics.washington.edu/phylip.html]
376. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees**. *Mol Biol Evol* 1987, **4**(4):406-425.
377. Jones D.T, Taylor W.R, Thornton J.M: **The rapid generation of mutation data matrices from protein sequences**. *Computer Applications in the Biosciences* 1992, **8**:275-282.
378. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0**. *Molecular Biology and Evolution* 2007, **24**:1596-1599.
379. Nei M., Kumar S: **Molecular Evolution and Phylogenetics**. New York: Oxford University Press; 2000
380. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0.** In.; 2004.
381. Lew-Tabor A, Rodriguez-Valle M, Moolhuijzen P. M, Bruyeres A. G: **Screening of Anti-Peptide Antibodies in vitro To Identify Potential Cattle Tick Vaccine Antigens**. In: *ICOPA XII: 2010; Melbourne, Australia*; 2010.
382. Bellgard MI, Guerrero F.D., Moolhuijzen PM, Schibeci D, Hunter A, Rodriguez-Valle M, Barrero R, Gondro C, Lew-Tabor AE: **Toward a genome sequence for *Rhipicephalus (Boophilus) microplus*: CattleTickBase available resources for the research community**. In.
383. Guerrero FD, Nene VM, George JE, Barker SC, Willadsen P: **Sequencing a new target genome: the Boophilus microplus (Acari: Ixodidae) genome project**. *J Med Entomol* 2006, **43**(1):9-16.
384. Lamoureux D, Peterson DG, Li W, Fellers JP, Gill BS: **The efficacy of Cot-based gene enrichment in wheat (Triticum aestivum L.)**. *Genome* 2005, **48**(6):1120-1126.
385. Glover DM, Kidd SJ, Roiha HT, Jordan BR, Endow S, Appels R: **Interrupter Sequences that are widely distributed in the *Drosophila* genome**. *Biochemical society transcations* 1978, **6**:732-736.
386. Wiener L, Schuler D, Brimacombe R: **Protein binding sites on Eschenichia coli 16S ribosomal RNA; RNA regions that are protected by proteins S7, S9 and S19, and by proteins S8, S15 and S17**. *Nucleic Acids Research* 1988, **16**(4):1233-1250.
387. Wiener L, Schuler D, Brimacombe R: **Protein binding sites on Eschenichia coli 16S ribosomal RNA; RNA regions that are protected by proteins S7, S9 and S19, and by proteins S8, S15 and S17**. *Nucleic Acids Research* 1988, **16**(4):1233-1250.
388. Lagesen K, Hallin P, R√∏dland EA, St√¶rfeldt HH, Rognes T, Ussery DW: **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. *Nucleic Acids Res* 2007, **35**(9):3100-3108.
389. Bunikis J, Barbour AG: **Ticks have R2 retrotransposons but not the consensus transposon target site of other arthropods**. *Insect Mol Biol* 2005, **14**(5):465-474.

390. Untalan PM, Guerrero FD, Haines LR, Pearson TW: **Proteome analysis of abundantly expressed proteins from unfed larvae of the cattle tick, Boophilus microplus**. *Insect Biochem Mol Biol* 2005, **35**(2):141-151.
391. **iMOL** [http://www.pirx.com/iMol/]
392. Kumar A: **An Overview of Nested Genes in Eukaryotic Genomes**. *EUKARYOTIC CELL* 2009, **8**(9):1321–1329.
393. Campbell AG, Fessler LI, Salo T, Fessler JH: **Papilin: a Drosophila proteoglycan-like sulfated glycoprotein from basement membranes**. *J Biol Chem* 1987, **262**(36):17605-17612.
394. Fessler JH, Kramerova I, Kramerov A, Chen Y, Fessler LI: **Papilin, a novel component of basement membranes, in relation to ADAMTS metalloproteases and ECM development**. *Int J Biochem Cell Biol* 2004, **36**(6):1079-1084.
395. Kramerova IA, Kawaguchi N, Fessler LI, Nelson RE, Chen Y, Kramerov AA, Kusche-Gullberg M, Kramer JM, Ackley BD, Sieron AL *et al*: **Papilin in development; a pericellular protein with a homology to the ADAMTS metalloproteinases**. *Development* 2000, **127**(24):5475-5485.
396. Kramerova IA, Kramerov AA, Fessler JH: **Alternative splicing of papilin and the diversity of Drosophila extracellular matrix during embryonic morphogenesis**. *Dev Dyn* 2003, **226**(4):634-642.
397. Corral-Rodriguez MA, Macedo-Ribeiro S, Barbosa Pereira PJ, Fuentes-Prior P: **Tick-derived Kunitz-type inhibitors as antihemostatic factors**. *Insect Biochem Mol Biol* 2009.
398. Ribeiro JM, Makoul GT, Levine J, Robinson DR, Spielman A: **Antihemostatic, antiinflammatory, and immunosuppressive properties of the saliva of a tick, Ixodes dammini**. *J Exp Med* 1985, **161**(2):332-344.
399. Simpson A.J., Maxwell A.I., Govan J.R., Haslett C., J.M.. S: **Elafin (elastase-specific inhibitor) has anti-microbial activity against gram-positive and gram-negative respiratory pathogens**. *FEBS Lett* 1999, **452**(3):309-313.
400. Smith CD, Shu S, Mungall CJ, Karpen GH: **The Release 5.1 annotation of Drosophila melanogaster heterochromatin**. *Science* 2007, **316**(5831):1586-1591.
401. Matz MV: **Amplification of representative cDNA samples from microscopic amounts of invertebrate tissue to search for new genes**. *Methods Mol Biol* 2002, **183**:3-18.
402. Melen GJ, Pesce CG, Rossi MS, Kornblihtt AR: **Novel processing in a mammalian nuclear 28S pre-rRNA: tissue-specific elimination of an 'intron' bearing a hidden break site**. *EMBO J* 1999, **18**(11):3107-3118.
403. Anderson JM, Sonenshine DE, Valenzuela JG: **Exploring the mialome of ticks: an annotated catalogue of midgut transcripts from the hard tick, Dermacentor variabilis (Acari: Ixodidae)**. *BMC Genomics* 2008, **9**:552.
404. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM *et al*: **The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective**. *Genome Biol* 2002, **3**(12):RESEARCH0084.
405. Tu Z: **Genomic and evolutionary analysis of Feilai, a diverse family of highly reiterated SINEs in the yellow fever mosquito, Aedes aegypti**. *Mol Biol Evol* 1999, **16**(6):760-772.

406. Tu Z, Li S, Mao C: **The changing tails of a novel short interspersed element in Aedes aegypti: genomic evidence for slippage retrotransposition and the relationship between 3' tandem repeats and the poly(dA) tail**. *Genetics* 2004, **168**(4):2037-2047.
407. Nene V, Lee D, Quackenbush J, Skilton R, Mwaura S, Gardner MJ, Bishop R: **AvGI, an index of genes transcribed in the salivary glands of the ixodid tick Amblyomma variegatum**. *Int J Parasitol* 2002, **32**(12):1447-1456.
408. Nene V, Lee D, Kang'a S, Skilton R, Shah T, de Villiers E, Mwaura S, Taylor D, Quackenbush J, Bishop R: **Genes transcribed in the salivary glands of female Rhipicephalus appendiculatus ticks infected with Theileria parva**. *Insect Biochem Mol Biol* 2004, **34**(10):1117-1128.